Flood contributing factors analysis in Sacramento Valley, CA

Yun-Rou Lin
Department of Civil and Environmental Engineering (CEE)
University of Illinois Urbana-Champaign
Urbana, IL, USA
yunroul2@illinois.edu

Xueming Xu
Department of Civil and Environmental Engineering (CEE)
University of Illinois Urbana-Champaign
Urbana, IL, USA
xx37@illinois.edu

Jim Liu
Department of Civil and Environmental Engineering (CEE)
University of Illinois Urbana-Champaign
Urbana, IL, USA
jiml2@illinois.edu

KP Pane
Department of Civil and Environmental Engineering (CEE)
University of Illinois Urbana-Champaign
Urbana, IL, USA
kpane2@illinois.edu

I. Introduction

Flooding in California's Sacramento Valley poses significant risks to communities, economic activity, and transportation systems. Our project will integrate multiple datasets —including precipitation, distance to waterways, DEM, and land cover—to identify the parameters most critical in driving flood events. By examining two major flooding events in 2018 with these datasets, we will conduct sensitivity analysis to quantify the relative contribution of each factor.

The goal of this analysis is to develop a machine learning (ML) framework for flood prediction. By reducing the dimensionality of input data, we aim to improve both model interpretability and computational efficiency. Furthermore, this framework will enable scenario simulations—such as intensified rainfall or land cover changes—to assess potential shifts in flood risk under future conditions.

A. Dataset Description

Dataset	Source	Format	Description
Flood Data	[1]	TIFF	Spatial distribution of flooding area for two flood events in 2018. Pixel values indicate flood (1) and non-flood (0) area.
Precipitation	[2]	CSV	Fields: Date, Precipitation (mm), Daily precipitation during 2018-03-21 through 2018-03-23 and 2018-12-05 through 2018-12-09.
Precipitation	[3]	TXT	Fields: Monitoring site ID, date, time, precipitation (inches). Recorded precipitation depth during the 15-min interval during two flooding events in Sacramento Valley in 2018.
Distance to Waterway	[4]	SHP	Fields: Name, Type, ShapeLength, ShapeArea. Using GIS to filter rivers/streams and drawing the centerlines where distance grids can be computed.
DEM (Digital Elevation Model)	[5]	TIFF	Elevation around Sacramento Valley area.

Dataset	Source	Format	Description
Land Use / Land Cover data	[6]	TIFF	2018 National Land Cover Database (NLCD) raster data with 30-meter resolution. Fields: Land Cover Classification, Fraction Impervious Surface

II. Exploratory Data Analysis

Our study integrates multi-source geospatial and hydrometeorological data for the Sacramento Valley, including DEM (elevation), slope, land use/land cover, distance to waterways, precipitation, and flood extent for a December 2018 event. We quantify spatial relationships between flood labels and each factor using thematic maps and summary statistics. All layers are projected to the WGS 1984 UTM Zone 10N coordinate system.

A. Feature Selection

The motivation for which parameters to include in this study stemmed from existing research on flood susceptibility, namely from an article highlighting an ML approach to flooding in Morocco [7]. It specifies an array of independent variables that served as possible parameters for this project. Our team chose a set of relevant features by considering those variables in conjunction with officially recognized flood hazards. Hazards refer to environmental factors that can make an area more prone to experience flooding.

Resources for the Future identifies some examples of hazards to include rainfall patterns and frequency, elevation, and rivers [8]. Precipitation amount, elevation, and distance to waterways account for these hazards. The National Weather Service also points out that during river flooding, low-lying areas surrounding rivers are the first to be impacted [9]. This confirms the significance of elevation and distance to waterways, especially since the area of study focuses on the (locally) low-lying Sacramento Valley intersected by the Sacramento River.

It's rather intuitive that slope impacts flood risk, but existing research articles can be cited with this finding as well. An article on urban flood susceptibility in Mumbai states that low-slope areas facilitate much slower runoff than steeper areas, causing water to distribute slowly and potentially build up. They also noted that low-elevation areas tend to consist of more gradually-sloped land [10]. All that considered, slope was an important parameter for us to include.

The National Weather Service also highlights that urban areas are susceptible to flash flooding due to an increased presence of impenetrable surfaces, such as concrete, that inhibit water drainage into the soil. Additionally, they state that rocky and clay-like ground has poor drainage, which can increase flood likelihood [9]. These facts present the necessity to consider physical characteristics of the land, for which land cover is a great option for encompassing both natural and man-made classification types.

B. DEM (elevation)

DEM data for the study area were obtained from the USGS for a time period close to the December 2018 flooding event. The DEM was imported into ArcGIS Pro and re-projected to WGS 1984 UTM Zone 10N. Elevation and slope rasters were then derived from this dataset, and an appropriate color ramp was selected for visualization. The DEM indicates that elevation in the study area ranges from approximately −32 to 1,249 m above sea level. The mean elevation is about 122 m, which is much closer to the minimum than to the maximum value. This confirms that low-elevation terrain dominates the Sacramento Valley, consistent with regional topography and with our expectation that flooding is more likely to occur in local elevation minima.

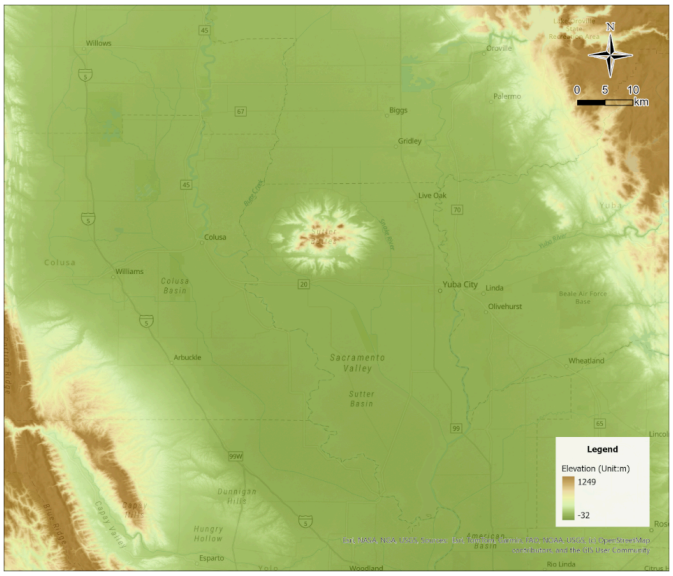

Fig. 1. DEM Data in Sacramento Valley, CA

C. Slope

From DEM, slope data was also derived in ArcGIS Pro. These results are shown in the figure below. Gentler slopes favor water accumulation and longer inundation residence times, while steep slopes promote rapid runoff. Slope consistently ranks among the most important predictors in susceptibility models. It can be observed that the slope values at the base of the valley are very low (close to 0 degrees), which is consistent with our expectation for flooding to occur in low-slope areas.

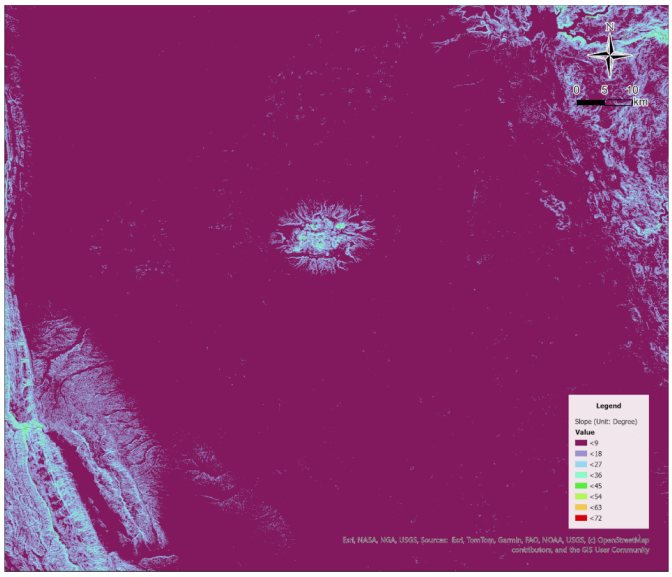

Fig. 2. Slope Data in Sacramento Valley, CA

D. *Aspect*

Aspect can be derived directly from the DEM, and the output represents the slope direction (e.g., N, NE, E). This factor influences landform development and helps explain how water may flow or accumulate during flooding events. In this study, we classified aspect into eight categories: flat (no slope), north (315°–22.5°), northeast (22.5°–67.5°), east (67.5°–112.5°), southeast (112.5°–157.5°), south (157.5°–202.5°), southwest (202.5°–247.5°), west (247.5°–292.5°), and northwest (292.5°–315°).

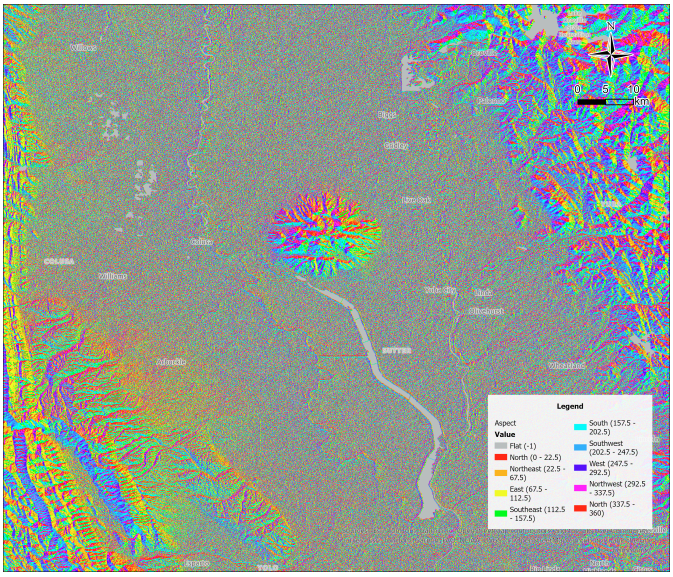

Fig. 3. Aspect Data in Sacramento Valley, CA

E. *Curvature*

Curvature describes how water flows across the surface, and it can also be derived from a DEM using ArcGIS. After computing this factor, a value of zero indicates areas with a higher potential flood risk. In this study, we classified plan curvature into three categories: concave (positive values), flat (zero values), and convex (negative values).

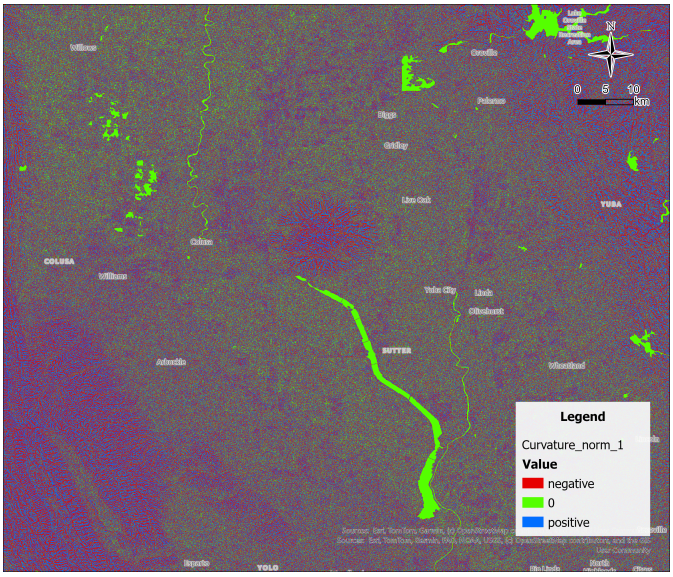

Fig. 4. Curvature Data in Sacramento Valley, CA

F. *Topographic Wetness Index (TWI)*

Previous studies have shown that the topographic wetness index (TWI) is closely related to flood occurrence. TWI represents how much moisture tends to accumulate in a watershed under the influence of gravity. TWI is calculated using the formula: TWI = ln(a / tanβ), where 'a' is the upslope contributing area per unit contour length, and 'β' is the local slope angle. Higher TWI values indicate areas more prone to saturation and flooding. In this study, we derived TWI from the DEM using ArcGIS Pro.

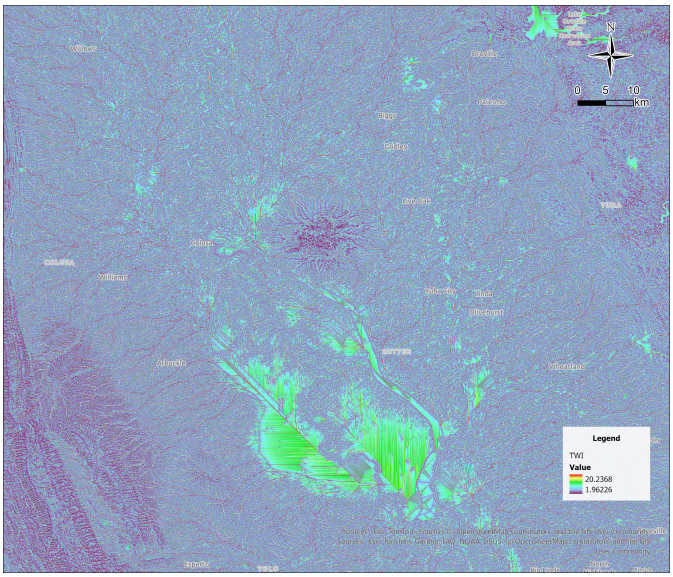

Fig. 5. TWI Data in Sacramento Valley, CA

G. Distance to waterway

Distance to waterways was computed from land cover data. We first extracted waterbodies from the land cover map to create a separate "open water" layer. Non-overlapping buffer rings were then generated around these waterbodies at specified distance intervals (in kilometers). Using the Intersect tool, we clipped flood polygons with each buffer ring so that distance class was stored as an attribute. For each ring, we calculated (i) the share of total flooded area and (ii) a normalized flood rate defined as (flooded area within a ring) / (total ring area).

Key processing steps included:

1) Erase (flood polygons minus permanent water);
2) Multiple Ring Buffer to create non-overlapping distance classes;
3) Intersect (flood polygons with buffer rings);
4) Calculate Geometry Attributes (area in km²);
5) Summary Statistics aggregated by distance class.

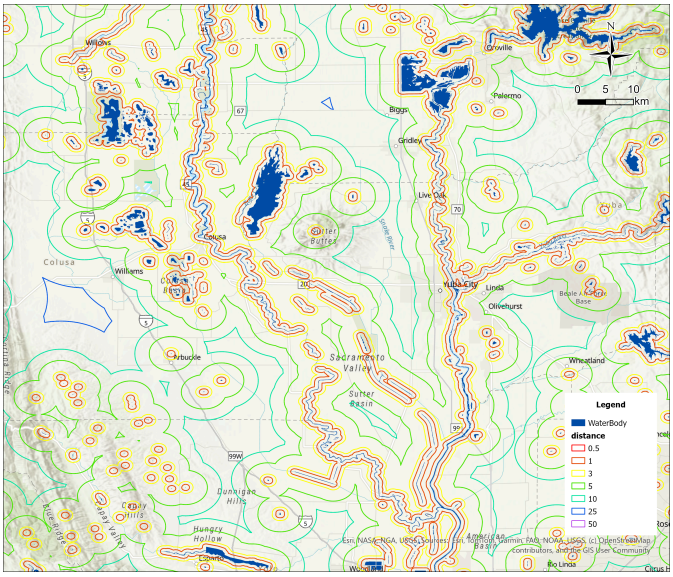

Fig. 6. Waterway Distance Buffer Rings

Within a 3 km corridor around waterways, normalized flood rates are relatively uniform: approximately 2.23–2.84% of each ring's area is flooded. Raw area shares are largest in the 1–3 km ring because that ring covers the largest land area. The near-uniform normalized rates suggest that, within 3 km of waterways, proximity alone does not fully control flood occurrence. Instead, topography (low basin slopes), local storage, and floodplain width likely govern where water spreads, which is consistent with our basin-scale hypothesis.

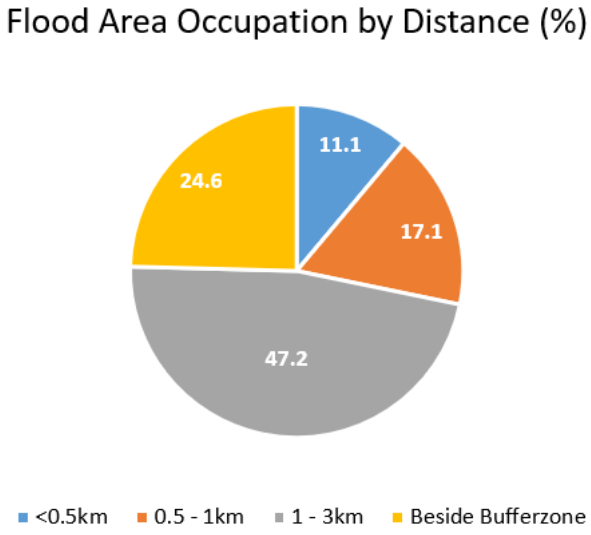

Fig. 7. Flood area by distance-to-water classes.

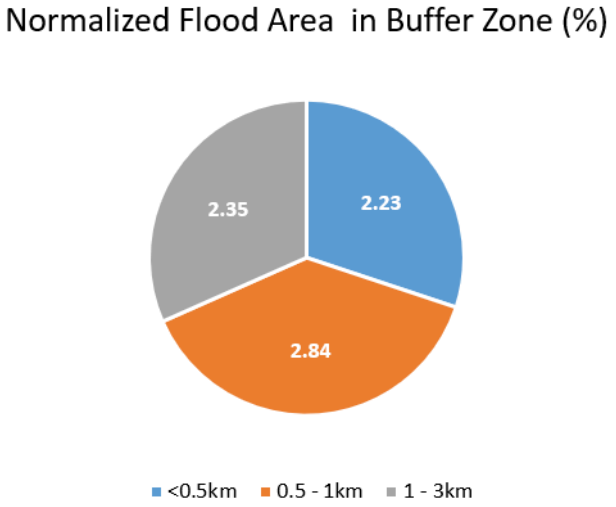

Fig. 8. Normalized flood rate (% of ring)

H. Land use/Land Cover

This dataset comes from the USGS National Land Cover Database (NLCD) from a 2018 dataset, which provides 30 meter resolution land cover classifications for the United States. The dataset classifies land cover into 16 different classes based on satellite imagery and other ancillary data. With interest in analyzing the impacts that land cover category has on flooding outcomes, a pie chart was created to visualize the distribution of land cover types among the total flooded area.

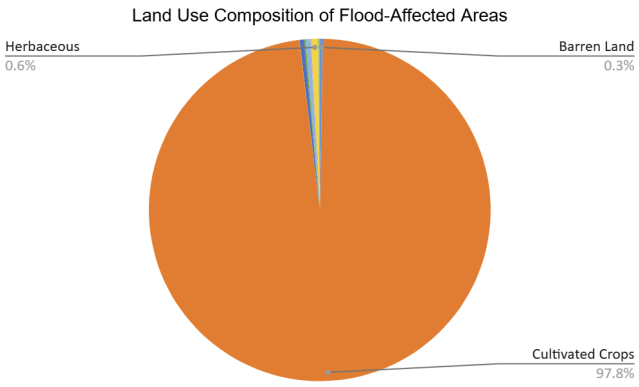

Fig. 9. Portion of Flooded Area Within Each Land Cover Type

To truly analyze the relationship between land cover and flooding, it is important to consider not just the proportion of each land cover type within the flooded area, but also the overall distribution of land cover types across the entire study area. This would allow for a more accurate assessment of whether certain land cover types are disproportionately represented in flooded areas compared to their prevalence in the landscape as a whole. As such, the percentage of flood prevalence was calculated within each individual land cover type, which provides a clearer picture of how likely each land cover type is to experience flooding.

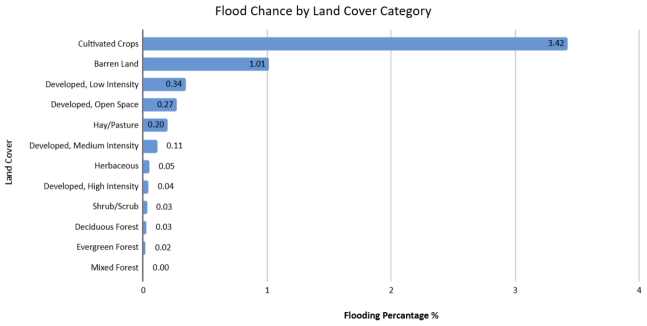

Fig. 10. Percentage of Area Flooded by Land Cover Type

Crop land overwhelmingly dominates the flooded area, with more than 3 times the rate of flooding as its runner up, barren land. This observation can likely be attributed to the fact that poor-drainage soils are preferable for agricultural activities due to the ability to retain moisture, but are consequently more prone to flooding. One surprising observation from this graph is that all developed land categories have a relatively low flood prevalence (below 1%). This could be due to the presence of stormwater management infrastructure in urban areas, such as storm drains and retention basins, which help mitigate flooding despite the high proportion of impervious surfaces. Other land cover types such as forest, shrub, and herbaceous also have low flood prevalence, likely due to their natural ability to absorb and slow down runoff.

I. *Daily/event precipitation*

Averaged Sacramento precipitation records for 2011–2024 show substantial interannual variability, with annual totals ranging from roughly 25 to 40 inches. Years such as 2017 and 2018 exhibit the highest totals and correspond to known regional flood events, whereas 2021–2022 are markedly drier and consistent with drought conditions. The frequency of rainfall events decreases as the intensity threshold increases. Light rain (≥ 0.01 in) occurs on approximately 40–70 days per year, moderate rain (≥ 0.10 in) on about 20–50 days per year, and heavy rain (≥ 1.00 in) on fewer than 6 days annually. Thus, most precipitation falls as frequent, low-intensity events, while a small number of high-intensity storms contribute disproportionately to flood potential. Extreme single-day precipitation maxima (1.5–3.5 in) closely track annual total rainfall, indicating that wetter years tend to experience both higher cumulative precipitation and more intense storms. Together, these patterns support the conclusion that short-duration, high-intensity rainfall events are key drivers of flooding risk in the Sacramento Valley.

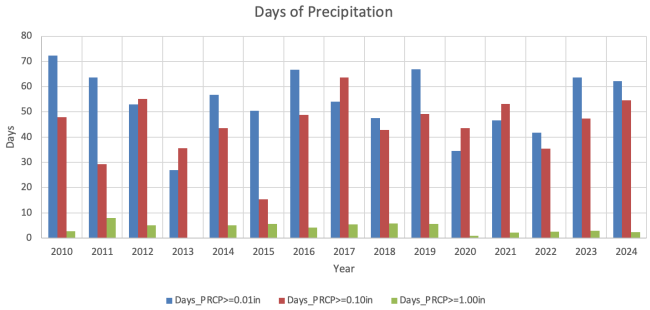

Fig. 11. Different Level of Precipitation by Days in Years

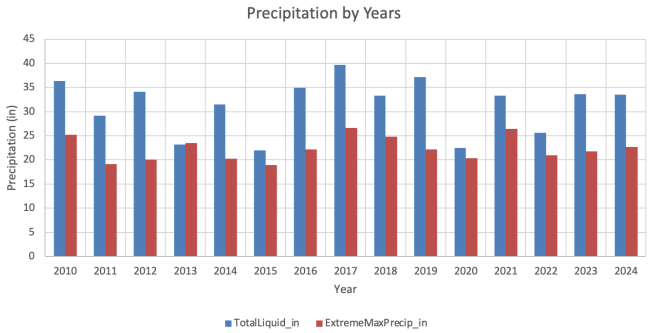

Fig. 12. Average Precipitation by Years

J. *Flood data*

Flood extents for the March 2018 and December 2018 events were imported into ArcGIS Pro and clipped to the study area. For March, the total flooded area is approximately 319 km², and for December, it is around 127 km², within a total area of interest of 11,835 km². Visual inspection of the flood map confirms that most inundation occurs in low-elevation zones at the base of the valley, consistent with our DEM and slope analyses. Because the observed flooding is concentrated within a relatively narrow elevation range, the

study may not fully capture the broader relationship between elevation and flood occurrence. Nonetheless, the satellite-derived flood-extent product provides pixel-level labels of flooded versus non-flooded areas for this event, which is suitable for training and validating data-driven flood susceptibility models.

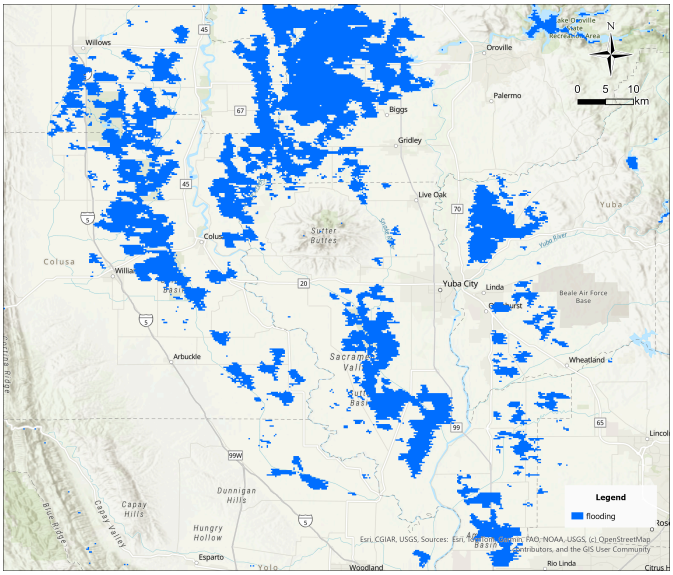

Fig. 13. Mar 2018 Flood Map Raster Data

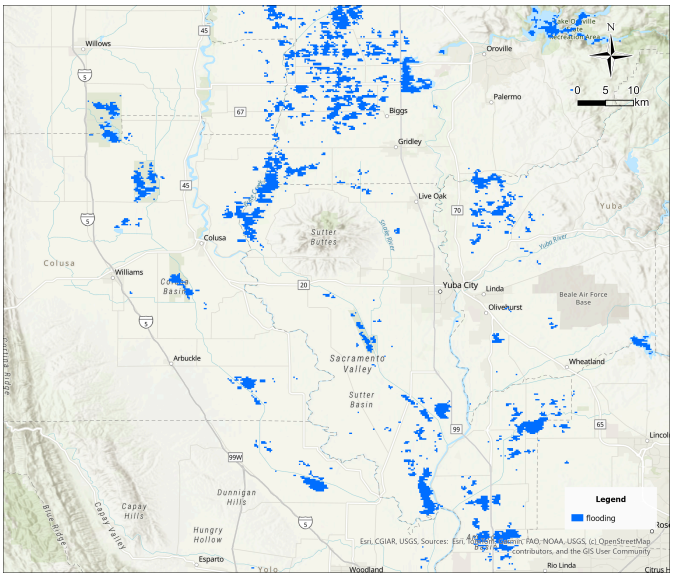

Fig. 14. Dec 2018 Flood Map Raster Data

III. Predictive Modeling

We proposed a supervised machine learning model with the following approach to identify the environmental variables most strongly associated with flood occurrence in Sacramento Valley. By conducting the sensitivity analysis, we are able to understand which factors are most influential which may improve the hazard mapping or the future assessment.

A. Objective

Objective Our goal is to identify the most influential environmental variables and hydrological factors that contribute the most to flood events in Sacramento Valley, CA.

B. Input features

Precipitation, elevation (DEM), slope, land cover type, and distance to waterways.

C. Model

According to Farhadi & Najafzadeh, Random Forest is a robust model for flood-related applications, especially when working with remote sensing data. It achieves high accuracy, reduces the risk of overfitting, and effectively captures nonlinear relationships among environmental variables [11]. Furthermore, Random Forest offers fast computation and simpler parameter tuning than other ML methods such as neural networks. Moreover, Random Forest provides more stable performance on moderately sized datasets. In addition, the Scikit-Learn implementation of Random Forest provides direct estimates of feature importance, which supports the objective of this project. As a result, Random Forest is a suitable method for our project.

For the loss function, the random forest also applies Gini impurity when doing the classification. Each tree in the forest relies on Gini to decide how to split the data. Since a random forest is essentially many decision trees trained on different subsets of the data and features, combining the predictions from all these trees makes the final result more robust and less sensitive compared to using a single decision tree.

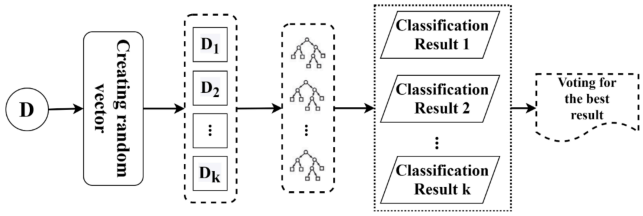

Fig. 15. Random Forest Structure [11]

D. Training and Validation

In this project, we wrote a script that reads multiple GIS raster layers and converts them into a clean pixel-level dataset. It then stacks the December and March event data and outputs a single CSV file (flood_data.csv) containing all input features and flood labels for modeling. Since our input classes are imbalanced, we addressed this issue by applying class_weight = {0: 1, 1: 6} to improve the model's ability to detect flooded pixels. For feature preparation, all geospatial layers were already aligned and extracted into a consistent pixel-level dataset, so no additional normalization was required for the Random Forest model.

E. Model Optimzation

We plan to adjust the hyperparameter to optimize the model. Our model:

```
rf = RandomForestClassifier(
n_estimators=200,
max_depth= 25,
min_samples_leaf = 3,
max_features="sqrt",
n_jobs=-1,
random_state=1234,
class_weight= {0: 1, 1: 6} )
```

a) *N_estimators:* represent the number of trees in the forest. A larger number of trees generally makes the model more stable, but it also increases training time. Since our dataset is moderately sized, we selected 200 trees. If the model accuracy does not improve after a certain point, there is no need to add more trees.

b) *Max_depth:* controls how deep each tree can grow. Since deeper trees can capture more complex patterns but also risk overfitting, we set an initial max_depth of 25. This allows the trees to learn detailed relationships without becoming too complex.

c) *Min_samples_leaf:* specifies the minimum number of samples required to be at a leaf node. Setting this to 3 helps prevent the model from creating leaves that are too specific to the training data, which can lead to overfitting. It encourages the model to generalize better by ensuring that each leaf has enough data points.

d) *Max_features:* determines how many features each tree can use when splitting. The default setting is "sqrt", which means each split considers √(number of input features). This helps introduce randomness and reduces overfitting.

e) *N_jobs = –1:* means the model will use all available CPU cores to speed up training.

f) *Random_state:* sets the random seed to ensure that the model results are reproducible.

g) *Class_weight:* addresses the class imbalance in our dataset by assigning a higher weight to the minority class (flooded pixels). This helps the model pay more attention to flooded areas during training, improving its ability to detect them.

h) *Optimization metric:* The metric used during training for the RandomForestClassifier is Gini Impurity. The post-training evaluation metrics are model accuracy, precision, recall, and F1 score.

F. Model Performance Evaluation

To evaluate te success of our predictive model, we use several standard performance metrics including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's ability to correctly classify flooded and non-flooded areas.

TABLE I
CONFUSION MATRIX

	True (Ground truth): Flooded	**True (Ground truth): Non-flooded**
Model prediction: Flooded	**1,674,768**	**347,078**
Model prediction: Non-flooded	**79,075**	**1,861,111**

The confusion matrix in Table 1 compares the model's flood prediction to the actual flooded areas observed in the real world (ground truth). True, or correct, predictions are highlighted in green, and false ones in red. "True (Model prediction)" represents pixels where the model predicted flooding, and "True (Ground truth)" represents the pixels that were actually flooded in the real world flood dataset. The model correctly predicted 1,674,768 pixels as flooded (true positive), but incorrectly predicted that an additional 347,078 pixels were flooded, which were not flooded in real life (false positive). 79,075 ground truth flood points were incorrectly predicted to be non-flooded (false negative), while the other 1861111 pixels that the model predicted as non-flooded were correct (true negative). This table makes it clear how often the model successfully detected true flooding versus how often it failed, and it provides a foundation for understanding the model's strengths and limitations.

TABLE II
TABLE 2. MODEL PERFORMANCE

Metric	**Value**	**Interpretation**
Accuracy	0.814	The accuracy remains high
Precision	0.349	Of the model's predicted flood area, 34.9% was truly flooded.
Recall	0.702	The model correctly found 70.2% of the flooded pixels.
F1-score	0.466	The moderate F1 score reflects the trade-off between relatively good recall and low precision.

With an accuracy of 81.4%, the model may appear reasonably successful at first glance, but the other metrics reveal important limitations. The model has a recall of 70.2%, meaning that it correctly identified a majority of the real-world flooded zones. However, the relatively low precision of 34.9% indicates that many of the pixels the model predicted as flooded were actually non-flooded. In other words, the model tends to over-predict flooding, which may be related to how it is learning from the imbalanced dataset.

The combination of high accuracy and low precision suggests that the dataset is still dominated by non-flooded pixels, and the accuracy is influenced strongly by correct non-flooded predictions. While recall shows that the model is sensitive to flooded areas, the low precision means that many non-flooded pixels are incorrectly labeled as flooded. The F1-score of 0.466 reflects this tradeoff between relatively high recall and low precision, indicating that there is room for improvement in balancing missed floods versus false alarms.

The precision represents how much of the model's predicted flood zones were truly flooded in the real-world data. With 34.9% precision, only about one-third of the predicted flooded area was actually flooded, so increasing this value would make the model more reliable for practical applications. Some methods for improving the model's performance include adding more data on flooded areas, rebalancing the dataset (for example, through oversampling or other techniques), and considering additional parameters that may affect flood risk. Adjusting the model's decision threshold

or using a weighting scheme that emphasizes more accurate flood predictions could also help reduce false positives and improve overall performance.

TABLE III
Random Forest feature importance

Factor	Importance
dem	0.245
slope	0.061
dist2river	0.037
landcover	0.213
aspect	0.074
curvature	0.063
TWI	0.050
rain	0.257

DEM (elevation) and precipitation are the two most important factors influencing flood occurrence, followed by land cover type. Slope, aspect, curvature, and topographic wetness index (TWI) have lower importance scores, indicating they contribute less to the model's predictions. These results align with our expectations, as elevation and rainfall are primary drivers of flooding, while land cover affects infiltration and runoff. The lesser importance of slope and other topographic features suggests that, in this region, they play a secondary role compared to elevation and precipitation.

IV. Discussion

Our results show that flooding in Sacramento Valley is mainly controlled by the interaction of elevation, land cover, and precipitation, with distance to waterways and detailed topographic metrics playing secondary roles. The December 2018 flood is concentrated in low-elevation, low-slope areas along the valley floor, while surrounding uplands remain largely unflooded. Within the 0–3 km river corridor, normalized flood rates are similar across distance bands, suggesting that once an area is "near" a river, local micro-topography and storage depressions are more important than exact distance to the channel. Land cover further modifies this pattern: cultivated crops have the highest flood prevalence, reflecting flat fields and limited drainage, whereas developed and vegetated areas show lower flood prevalence, likely due to stormwater infrastructure and higher infiltration and roughness.

The precipitation analysis supports the role of short-duration, high-intensity storms as the main trigger for flooding. Only a small fraction of days experience heavy rainfall, but these events contribute disproportionately to annual totals and coincide with known flood years such as 2018.

The Random Forest model captures part of this physical understanding but is clearly affected by class imbalance. Because the dataset contains far more non-flooded than flooded pixels, the model can achieve high overall accuracy by favoring non-flooded predictions, which leads to low recall for the flooded class. In other words, it is conservative and tends to miss many flooded pixels, limiting its usefulness for detecting all flooded areas. DEM (elevation) and precipitation are the two most important factors influencing flood occurrence, followed by land cover type. Slope, aspect, curvature, and topographic wetness index (TWI) have lower importance scores, indicating they contribute less to the model's predictions. These results align with our expectations, as elevation and rainfall are primary drivers of flooding, while land cover affects infiltration and runoff. The lesser importance of slope and other topographic features suggests that, in this region, they play a secondary role compared to elevation and precipitation, and the low importance of distance to river is consistent with our finding that proximity to channels within 3 km does not sharply separate flooded from non-flooded pixels.

This study has several limitations. The model is trained on 2 flood events and uses mostly static predictors at 30 m resolution, so event-specific patterns, unresolved small-scale features, and missing dynamic variables (e.g., soil moisture, river stage, reservoir operations) all restrict generalization. In future work, rebalancing the training data (e.g., oversampling flooded pixels), tuning the classification threshold, and evaluating performance with precision–recall metrics could better handle class imbalance. Incorporating multiple flood events, additional hydrologic variables, and models that account for spatial context would likely improve flood detection and make this framework more useful for operational flood-risk assessment.

V. Sources

[1] Cloud to Street & Dartmouth Flood Observatory, "Global Flood Database."

[2] National Centers for Environmental Information (NCEI), NOAA, "Climate Data Online (CDO)."

[3] U.S. Geological Survey, "National Water Information System: Real-Time Water Data."

[4] Esri & U.S. Geological Survey, "USA Detailed Water Bodies." 2023.

[5] U.S. Geological Survey, "The National Map Data Download Application."

[6] Multi-Resolution Land Characteristics Consortium (MRLC), "MRLC Data Portal."

[7] S. Hitouri *et al.*, "Flood Susceptibility Mapping Using SAR Data and Machine Learning Algorithms in a Small Watershed in Northwestern Morocco," *Remote Sensing*, vol. 16, no. 5, 2024, doi: 10.3390/rs16050858.

[8] M. A. Wall, S. Pesek, and D. Peterson, "Flooding in the United States 101: Causes, Trends, and Impacts," *Resources Magazine*, Sept. 2023, [Online]. Available: https://www.rff.org/publications/explainers/flooding-in-the-united-states-101-causes-trends-and-impacts/

[9] National Weather Service, U.S. Department of Commerce, "Flood Related Hazards."

[10] R. Veerappan and S. Sayed, "Urban flood susceptibility zonation mapping using evidential belief function, frequency ratio and fuzzy gamma operator models in GIS: a case study of Greater Mumbai, Maharashtra, India," *Geocarto International*, Mar. 2020, doi: 10.1080/10106049.2020.1730448.

[11] H. Farhadi and M. Najafzadeh, "Flood Risk Mapping by Remote Sensing Data and Random Forest Technique," *Water*, vol. 13, p. 3115, 2021, doi: 10.3390/w13213115.