Geospatial Machine Learning Framework for Risk-Adjusted Watermelon Farming Optimization in the Gandak Basin

I. Foundational Data Acquisition and Google Colab Environment Setup

The development of a robust Machine Learning (ML) framework for agricultural risk analysis begins with establishing a stable cloud environment and securing high-resolution, geospatially referenced data. Google Colaboratory (Colab) provides a convenient, no-setup Linux environment capable of running Python logic and Bash commands, making it ideal for processing geospatial workflows.[1]

I.A. Google Colab Environment Setup and Library Installation

Implementing hydrological and spatial analysis within Colab necessitates the installation of a specialized geospatial software stack. Because Colab runs in a cloud virtual machine, system dependencies, which often involve complex C-libraries (GEOS, GDAL, PROJ), must be managed explicitly.[3]

The core Python libraries required for raster and vector data manipulation are rasterio and geopandas.[4] While conda is generally the recommended manager for this stack on local machines [3], Colab environments typically require using !pip install commands to load these packages and their dependencies.

For advanced hydrological feature engineering, the workflow necessitates specialized tools.

Calculating the Topographic Wetness Index (TWI) and the Terrain Ruggedness Index (TRI) requires sophisticated raster processing capabilities.[7] Packages such as xdem offer efficient methods for calculating DEM derivatives like TRI, which is defined by the square root of squared differences with neighboring pixels.[8] For large, high-resolution datasets, if localized processing within the standard Colab runtime proves inefficient, the preliminary feature calculation can be shifted to highly parallelized cloud platforms like Google Earth Engine (GEE), which is available for non-commercial use and compatible with Colab interfaces.[9] This strategic shift significantly optimizes the workflow by downloading simplified feature arrays rather than the raw, massive Digital Elevation Model (DEM).

I.B. High-Resolution Digital Elevation Model (DEM) Selection and Access

The precision of flood modeling is fundamentally dependent on the spatial resolution of the terrain data.

I.B.1. Primary DEM Source and Access Link

The highest recommended resolution for this application is the **12.5-meter dataset derived from the Advanced Land Observing Satellite (ALOS) Phased Array type L-band Synthetic Aperture Radar (PALSAR)**.[10] This resolution captures finer terrain details, which is crucial for improving the precision of flood behavior modeling, particularly when coupled with hydrodynamic models like HEC-RAS in regional studies of Bihar.[10]

While a direct, static download link is challenging to provide universally, access to the ALOS PALSAR DEM (12.5m) for the Bihar Gandak region is primarily available through the data archives of the Alaska Satellite Facility (ASF) Distributed Active Archive Center (DAAC) or JAXA's ALOS data portals. Users typically require a free registration and login (like a NASA Earthdata account) to select and download the specific GeoTIFF tile covering the lower Gandak basin.[12] This data acquisition step directly addresses the user request for the elevation data link.

I.B.2. Data Fidelity and Fallback Options

Studies of the flood-prone lower Gandak River basin utilize this high-resolution ALOS DEM for quantitative analysis of morphometric characteristics across sub-watersheds.[11] Lower-resolution alternatives, such as the 30-meter products from the Shuttle Radar Topography Mission (SRTM) or the ASTER Global DEM (GDEM version 3), are available via USGS EarthExplorer or AppEEARS.[12] However, the use of coarser data (e.g., 30m resolution) introduces a critical trade-off: these lower-resolution products tend to smooth terrain details.[10] This smoothing reduces the fidelity of derived hydrological features like Slope, TRI, and TWI, which are essential for accurate spatial risk classification. For a scientifically rigorous suitability map, securing the 12.5m DEM is strongly preferred.

I.C. Hydrological and Meteorological Data Pipeline Establishment

The ML framework requires historical data sets spanning river behavior (for risk calibration) and atmospheric conditions (for yield forecasting).

I.C.1. Gandak River Hydrology Data

The authoritative source for historical flood data in the region is the Flood Management Improvement Support Centre (FMISC) of the Water Resources Department, Government of Bihar.[14] This center provides access to Central Water Commission (CWC) gauge data, Daily Flood Bulletins, and Barrage Status Reports, including those for the Valmikinagar Barrage on the Gandak River.[14]

The critical feature to extract is the time series of the **Maximum Historical Water Level (HFL_{Max}) or Danger Level (DL)** from the closest CWC gauge serving the target farming area.[15] These measurements define the highest observed flood boundary, which serves as the physical anchor for the spatial suitability analysis. It must be noted that historical time series for the Burhi-Gandak and similar tributaries are often non-continuous, containing data predominantly for the monsoon period, generally from June 15th to October 15th.[16] Given this temporal discontinuity, the classification model cannot rely on real-time water level inputs. Instead, the highest recorded HFL_{Max} becomes a **static, binary constraint** (X_1), representing the worst-case historical flood hazard for spatial suitability mapping.

I.C.2. Meteorological and Agricultural Climate Data

- **Rainfall Data:** Flood hazard assessment is highly sensitive to precipitation.[17] Historical daily gridded rainfall data, ideally at 0.25° resolution, can be accessed from the Indian Meteorological Department (IMD) via their API services.[18] This dataset is necessary for generating the Maximum Historical Rainfall feature (X_5), which contributes to the Stage 1 flood susceptibility model.
- **Agricultural Variables:** Climate variables required for yield and profitability modeling (Stage 2) can be sourced efficiently from the NASA Prediction Of Worldwide Energy Resources (POWER) project.[20] The POWER Data Access Viewer (DAV) provides geospatially enabled solar and meteorological parameters, such as maximum temperature, relative humidity, and solar radiation, which are specifically formulated for assessing agricultural systems and integrating into custom software applications.[21]

II. Hydrological Risk Zoning and Feature Engineering

This phase transforms the acquired DEM and hydrological data into ML-ready features, culminating in the definition of Secure and High-Risk zones, fulfilling the second objective of the user query.

II.A. Calculation of the Elevation Margin Feature (X_1)

The foundational geospatial feature for flood risk modeling is the relative elevation of the land surface compared to the highest recorded flood level.[22]

1. **HFL Proxy Derivation:** The maximum water surface elevation (HFL_{Max}) established in Section I.C.1 is used as the absolute elevation threshold for historical inundation. This value is derived from the nearest Gandak River CWC gauge station, such as the Valmikinagar Barrage, which provides a physical reference point.[14]
2. Elevation Margin (EM) Calculation: The Elevation Margin (EM) is calculated per raster cell using the formula:

$$\text{EM} = \text{Local DEM Elevation} - \text{HFL}_{\text{Max}}$$

A positive EM indicates the land surface is elevated above the maximum observed flood level, suggesting relative safety. A negative EM indicates historical inundation occurred at that location.6

II.B. Definition of Risk Zones and Target Variable Generation

For commercial farming, particularly for crops sensitive to waterlogging like watermelon, land suitability is not solely determined by whether it stays dry during a flood, but also whether the root zone remains unsaturated.

II.B.1. Incorporating the Agricultural Safety Buffer (B_{Safety})

The analysis must extend beyond simple inundation. When the river water surface elevation (WSE) is high, even adjacent fields may experience crop failure due to the resulting rise in the groundwater table, which saturates the soil via capillary action.[23] Accurate estimation of capillary rise is vital for developing environmental strategies that prevent soil deterioration and ensure water is available for plant irrigation.[24] For typical Gangetic alluvial soils, a maximum height of capillary rise often requires establishing a minimum safety margin (B_{Safety}) above the HFL_{Max}, typically defined as 1.5 to 2.0 meters.

This requirement means that a field must not only be above the historical flood level but also sufficiently distant (in terms of elevation) from it to prevent root saturation. The model is thus built conservatively, prioritizing risk minimization necessary for sustainable investment validation.[25]

II.B.2. Risk Zoning Criteria (Query Point 2)

By integrating the Elevation Margin (EM) with the B_{Safety} buffer, the land is classified into three distinct categories, conceptually aligning with FEMA flood zone definitions for effective risk communication.[26] These zones define the binary target variable,

Y_{Safety}, for the Stage 1 classification model.

Flood Risk Zoning Criteria for Watermelon Farming

Risk Zone Category	Target Variable (YSafety)	Primary Criteria (Elevation Margin)	Risk Profile (FEMA Proxy)	ML Training Sample
High Risk (Unsuitable)	0 (Unsafe)	$\text{EM} \leq 0$	FEMA Zone A/AE (1% annual chance of inundation) [26]	Negative
Moderate Risk (Conditional)	0 (Unsafe)	$0 < \text{EM} \leq \text{B}_{\text{Safety}}$ (e.g., 1.5m)	FEMA Zone X (Shaded, moderate hazard, water table saturation risk)	Negative (Risk Avoidance)
Secure Zone (Highly Suitable)	1 (Safe)	$\text{EM} > \text{B}_{\text{Safety}}$	FEMA Zone C/X (Unshaded, minimal hazard/above 500-year flood level) [26]	Positive

By aggregating both inundated land and saturation-risk land into the $Y=0$ (Unsafe) category, the classification model is trained to aggressively avoid high-risk agricultural areas.

II.C. Advanced Topographic Feature Engineering (X_{2-6})

To enhance the performance of the ML model, complex features that quantify micro-drainage characteristics and terrain structure are derived from the high-resolution DEM.

1. **Topographic Wetness Index (TWI) (X_2):** TWI is a GIS-based indicator that

combines local slope and upstream contributing area to estimate how water accumulates.[7] It is widely used to identify potential soil moisture and saturation zones.[7] In flat alluvial plains like the Gandak basin, high TWI values strongly correlate with poor drainage, which is a significant secondary hazard leading to waterlogging and disease, even if the area is not directly inundated.[29] The quality of the TWI feature is highly dependent on the precision of the flow accumulation calculation, underscoring the necessity of using the high-resolution 12.5m DEM.[10]

2. **Terrain Ruggedness Index (TRI) (X_3):** TRI measures the complexity of the terrain by calculating the average variation in elevation relative to surrounding pixels.[8] While the Gangetic plain is generally flat, localized high TRI values indicate a rugged landscape likely to experience greater surface runoff and erosion during heavy rainfall.[29] This feature helps delineate areas of high-velocity flow accumulation versus areas of slow, ponding accumulation.
3. **Contextual and Meteorological Features (X_{4-6}):**
   - **Distance to River (X_4):** Measures the shortest Euclidean distance from each land cell to the main Gandak river channel. This feature serves as a proxy for flood exposure and the travel time of flood waves.[22]
   - **Maximum Historical Rainfall (X_5):** Incorporates the non-topographical driver of flood hazard, derived from gridded IMD data.[17] Analysis has shown that elevation and precipitation are the factors most significantly contributing to flood hazard assessment in these regions.[17]
   - **Soil and Land Use (X_6):** Categorical data representing the specific alluvial soil properties, which influence permeability and capillary action.[30]

III. Machine Learning Framework for Suitability and Profitability

The final stage involves constructing and training a dual-stage ML architecture to output the probability of safe farming and the conditional profit percentage (Query Point 3). This hybrid model leverages classification for risk analysis and regression for economic forecasting.[25]

III.A. Stage 1: Safe Farming Classification Model (P_{Safety})

The first model is trained to predict the probability (P_{Safety}) that a given farming location is suitable and safe from catastrophic flood events ($Y_{\text{Safety}}=1$).

1. **Model Selection and Training:** Ensemble methods consistently demonstrate superior predictive ability in flood susceptibility mapping.[32] **XGBoost Classifier** is selected as the primary modeling technique. XGBoost and similar boosting algorithms often achieve high Area Under the Curve (AUC) scores (e.g., >0.9) and are adept at capturing the complex, non-linear relationship between topographic features (EM, TWI, TRI) and the binary outcome of flood survival.[17]
2. **Rigor and Validation:** The training data uses the geographically distributed binary risk labels (0 or 1) defined in Section II.B.2. Validation emphasizes high *Specificity*—the model's ability to correctly identify and exclude truly unsafe or high-risk areas. Furthermore, advanced model interpretability methods, such as **SHAP or Boruta analysis**, are employed to verify that the classification is physically grounded, confirming that Elevation Margin (X_1) and Max Rainfall (X_5) are the statistically dominant factors driving the flood risk prediction.[17]

III.B. Stage 2: Watermelon Profitability Regression Model ($E[\text{Profit}]$)

The second model predicts the financial outcome, conditional on the physical safety established in Stage 1.

1. **Target Variable Definition:** The financial outcome is the **Expected Profit Percentage**, calculated using input variables such as yield, local market price, and cultivation costs.[33]
$$\text{Profit Percentage} = \frac{(\text{Yield} \times \text{Market Price}) - \text{Cultivation Cost}}{\text{Cultivation Cost}} \times 100$$
2. **Model Selection and Features:** A **Random Forest Regressor (RFR)** is highly suitable for agricultural yield prediction.[31] RFR is an ensemble learning technique that effectively handles a diverse set of inputs and reduces the risk of overfitting by averaging multiple decision trees.[35]
   - **Inputs:** The RFR uses historical data on climate parameters (TMax, Solar Radiation, RH from NASA POWER [20]), regional economic data (Cultivation Cost, Market Price [33]), and a critical derived feature: the predicted safety probability, P_{Safety}, outputted from Stage 1. Including P_{Safety} allows the regression model to learn localized yield suppressions that correlate with proximity to risk zones, even if inundation does not occur.

III.C. Integration: Final Conditional Expected Profit Probability

(P_{Final})

The final required metric synthesizes the physical risk and the economic reward into a single, actionable probability score.[25] This metric translates the raw expected profit into a value that accounts for the inherent probability of failure.

1. **Integration Formula:** The final risk-adjusted expected profit percentage (P_{Final}) is calculated by multiplying the probability of safety by the expected profit percentage under safe conditions:
$$P_{\text{Final}} = P_{\text{Safety}} \times E$$
2. **Actionability and Investment Profile:** This probabilistic weighting provides a quantitative metric for investment comparison. For instance, a site promising 90% profit but having only a 50% chance of surviving the monsoon season yields a risk-adjusted return of 45%. By comparing sites based on P_{Final}, agricultural investors can optimize resource allocation, select sites with the most favorable risk-reward balance, and promote more sustainable and resilient farming practices.[25] The resulting spatially explicit P_{Final} map directly addresses the user's need for a final probability outcome that integrates both suitability and profitability.

IV. Synthesis, Validation, and Strategic Recommendations

The implemented framework provides a robust spatial decision support tool for watermelon cultivation in the Gandak basin.

IV.A. Model Validation and Feature Interpretation

Rigor in ML modeling necessitates validating that the predictive success is driven by physically meaningful factors. Interpreting the classification model (Stage 1) through SHAP or Boruta analysis is essential to confirm that Elevation Margin (X_1), Topographic Wetness Index (X_2), and Maximum Historical Rainfall (X_5) are the predominant drivers of flood susceptibility prediction, thereby confirming the model's hydrological validity.[17]

The performance of the Stage 1 model should be measured using AUC and Specificity, emphasizing the confidence in correctly identifying high-risk areas. The Stage 2 regression model is evaluated using metrics like R^2 and Root Mean Square Error (RMSE), ensuring accurate prediction of yields and prices, particularly under varied climate scenarios derived from the NASA POWER data.

IV.B. Strategic Farming Recommendations

The final P_{Final} map is the basis for actionable operational strategies in the flood-prone environment of the Gandak region.

1. **Secure Zone Strategy (P_{Final} High):** Areas classified as Secure (EM exceeding the B_{Safety} buffer) are optimal for standard, high-yield watermelon cultivation. These zones should be prioritized for long-term agricultural investment, as the risk of catastrophic loss from historical flood levels or waterlogging is minimal.
2. **Moderate Zone Management (P_{Final} Medium):** Areas falling within the $0 < \text{EM} \leq \text{B}_{\text{Safety}}$ range require mitigation strategies.
    - **Timing Mitigation:** Given that the majority of significant flood events and high river levels occur during the monsoon window (June 15th to October 15th) [16], farmers can schedule planting to ensure the harvest is completed before the peak flood period arrives. The Stage 2 profitability model also provides insights into optimizing planting schedules to align harvest with the most profitable market months.[34]
    - **Engineering Mitigation:** For moderate zones characterized by high TWI values (indicating poor drainage and waterlogging risk), field modifications are recommended. This includes implementing specialized agricultural engineering solutions such as deep drainage ditches or permanent raised-bed farming systems, which physically lift the root zone further above the rising water table, effectively increasing the agricultural safety buffer (B_{Safety}).[37]
3. **Risk Avoidance:** Sites categorized as High Risk ($\text{EM} \leq 0$) should be excluded from commercial watermelon farming entirely due to the near-certainty of annual loss, regardless of expected profitability. The conservative classification model ensures that the ML output serves as a robust risk filtration mechanism for capital deployment.

Works cited

1. Google Colaboratory | ArcGIS API for Python - Esri Developer, accessed November 19, 2025, https://developers.arcgis.com/python/latest/guide/install-and-set-up/google-colab/
2. Get started with GRASS in Google Colab - Learn GRASS - OSGeo, accessed November 19, 2025,

https://grass-tutorials.osgeo.org/content/tutorials/get_started/grass_gis_in_google_colab.html
3. Installation — GeoPandas 1.1.1+0.ge9b58ce.dirty documentation, accessed November 19, 2025, https://geopandas.org/en/stable/getting_started/install.html
4. rasterio.ipynb - Colab, accessed November 19, 2025, https://colab.research.google.com/github/giswqs/geog-312/blob/main/book/geospatial/rasterio.ipynb
5. Reading Datasets — rasterio 1.4.3 documentation, accessed November 19, 2025, https://rasterio.readthedocs.io/en/stable/topics/reading.html
6. Simulating Flood Inundation with Python and Elevation Data: A Beginner's Guide, accessed November 19, 2025, https://towardsdatascience.com/simulating-flood-inundation-with-python-and-elevation-data-a-beginners-guide/
7. Layer: Topographic Wetness Index (ID: 6), accessed November 19, 2025, https://maps.lakecountyil.gov/arcgis/rest/services/GISMapping/WABDrainage/MapServer/6
8. Terrain attributes - xDEM - Read the Docs, accessed November 19, 2025, https://xdem.readthedocs.io/en/stable/terrain.html
9. Data Access - Landsat Science - NASA, accessed November 19, 2025, https://landsat.gsfc.nasa.gov/data/data-access/
10. (PDF) Potential Dam Breach Analysis and Flood Wave Risk Assessment Using HEC-RAS and Remote Sensing Data: A Multicriteria Approach - ResearchGate, accessed November 19, 2025, https://www.researchgate.net/publication/348929933_Potential_Dam_Breach_Analysis_and_Flood_Wave_Risk_Assessment_Using_HEC-RAS_and_Remote_Sensing_Data_A_Multicriteria_Approach
11. Assessment of flood hazards using morphometric compound factor and hypsometric integral in lower Gandak basin, India - ResearchGate, accessed November 19, 2025, https://www.researchgate.net/publication/395350758_Assessment_of_flood_hazards_using_morphometric_compound_factor_and_hypsometric_integral_in_lower_Gandak_basin_India
12. 30-Meter SRTM Elevation Data Downloader - Derek Watkins, accessed November 19, 2025, https://dwtkns.com/srtm30m/
13. Where can I get global elevation data? | U.S. Geological Survey - USGS.gov, accessed November 19, 2025, https://www.usgs.gov/faqs/where-can-i-get-global-elevation-data
14. FMISC WRD Bihar, accessed November 19, 2025, https://www.fmiscwrdbihar.gov.in/
15. FMISC WRD Bihar, accessed November 19, 2025, https://fmiscwrdbihar.gov.in/fmis/statistics.html
16. Operational Research to Support Mainstreaming of Integrated Flood Management under Climate Change - GOV.UK, accessed November 19, 2025, https://assets.publishing.service.gov.uk/media/57a0899bed915d622c0002e5/61183_ADB-MoU-Research-outputs_Vol5a-Modelling-Report-BG-Final.pdf

17. Integrating machine learning and geospatial data analysis for comprehensive flood hazard assessment - PMC - PubMed Central, accessed November 19, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11297827/
18. IMD Monthly Rainfall - IDS-DRR, accessed November 19, 2025, https://drr.open-contracting.in/en/datasets/2996df05-cc01-479a-8d03-1e8a3cd26fd1
19. List of API's of India Meteorological Department, accessed November 19, 2025, https://mausam.imd.gov.in/imd_latest/contents/api.pdf
20. Data Access Viewer (DAV) - NASA POWER, accessed November 19, 2025, https://power.larc.nasa.gov/data-access-viewer/
21. Data Access Viewer (DAV) - NASA POWER, accessed November 19, 2025, https://power.larc.nasa.gov/docs/tutorials/data-access-viewer/user-guide/
22. Flood risk assessment using machine learning, hydrodynamic modelling, and the analytic hierarchy process | Journal of Hydroinformatics | IWA Publishing, accessed November 19, 2025, https://iwaponline.com/jh/article/26/8/1852/103822/Flood-risk-assessment-using-machine-learning
23. WATER PRODUCTIVITY ASSESSMENT: Measuring and Mapping Methodologies Basin Focal Project Working Paper no. 2, accessed November 19, 2025, https://assets.publishing.service.gov.uk/media/57a08c45e5274a31e00010ea/AgriculturalWaterProductivityBFPwp02Draft03.pdf
24. An approach for quick estimation of maximum height of capillary rise - ResearchGate, accessed November 19, 2025, https://www.researchgate.net/publication/269777696_An_approach_for_quick_estimation_of_maximum_height_of_capillary_rise
25. Innovative Approaches to Agricultural Risk with Machine Learning - The Science and Information (SAI) Organization, accessed November 19, 2025, https://thesai.org/Downloads/Volume15No7/Paper_104-Innovative_Approaches_to_Agricultural_Risk.pdf
26. Definitions of FEMA Flood Zone Designations Moderate to Low Risk Areas High Risk Areas, accessed November 19, 2025, https://www.berkspa.gov/getmedia/9388a72b-3d90-40a0-9d71-3243718d8681/FEMA-Flood-Zone-Definitions.pdf
27. FEMA Flood Zone Definitions - City of Newburyport, accessed November 19, 2025, https://www.cityofnewburyport.com/planning-development/files/fema-flood-zone-definitions
28. Topographic Wetness Index (TWI) Modelling Using Monte Carlo Simulation in Python & PCRaster | by Julian Manning | Medium, accessed November 19, 2025, https://medium.com/@julian.manning/topographic-wetness-index-twi-modelling-using-monte-carlo-simulation-in-python-pcraster-ddd2f916fd81
29. Full article: Machine learning-based assessment of flood susceptibility in the Eastern Mediterranean: a case study of Baniyas River basin - Taylor & Francis Online, accessed November 19, 2025, https://www.tandfonline.com/doi/full/10.1080/19475705.2025.2524417

30. (PDF) Using machine learning for land suitability classification - ResearchGate, accessed November 19, 2025, https://www.researchgate.net/publication/282925748_Using_machine_learning_for_land_suitability_classification
31. Multimodal Machine Learning Based Crop Recommendation and Yield Prediction Model, accessed November 19, 2025, https://www.researchgate.net/publication/366762972_Multimodal_Machine_Learning_Based_Crop_Recommendation_and_Yield_Prediction_Model
32. Full article: SAR-driven flood inventory and multi-factor ensemble susceptibility modelling using machine learning frameworks, accessed November 19, 2025, https://www.tandfonline.com/doi/full/10.1080/19475705.2024.2409202
33. Best Profitable Crops Prediction with Profit, Cost, and Farmland Optimization using Machine Learning - MIR Labs, accessed November 19, 2025, https://www.mirlabs.org/ijcisim/regular_papers_2023/Paper49.pdf
34. Optimizing agricultural yield: a predictive model for profitable crop harvesting based on market dynamics - Frontiers, accessed November 19, 2025, https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1567333/full
35. Predicting land suitability for wheat and barley crops using machine learning techniques, accessed November 19, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12059010/
36. Investigating Flood Impact on Crop Production under a Comprehensive and Spatially Explicit Risk Evaluation Framework - ResearchGate, accessed November 19, 2025, https://www.researchgate.net/publication/359626541_Investigating_Flood_Impact_on_Crop_Production_under_a_Comprehensive_and_Spatially_Explicit_Risk_Evaluation_Framework
37. Crop Prediction Model Using Machine Learning Algorithms - MDPI, accessed November 19, 2025, https://www.mdpi.com/2076-3417/13/16/9288