The State of Machine Learning in Flood Monitoring and Prediction Utilizing Sentinel-1 and Sentinel-2 Data: A Systematic Literature Review

A. S. M Borhanul Islam[a,1], Sumaiya Tabassum[a], Md Asif Bin Khaled[b,1,*]

[a]*Department of Computer Science and Engineering, Independent University, Bangladesh,*
[b]*Center for Computational & Data Sciences, Independent University, Bangladesh,*

Abstract

Background: Floods are some of the most destructive natural disasters on Earth, and their occurrence and intensity are increasing due to climate change and human action. Traditional flood monitoring approaches have difficulties taking advantage of limited spatial-temporal coverage and operational lag times in emergencies. The application of next-generation remote sensing combined with machine learning (ML) and deep learning (DL) techniques, offers a robust mechanism for enhancing flood monitoring, prediction, and response.

Objective: A systematic review is presented to assess and synthesize studies using Sentinel-1 and Sentinel-2 data with ML/DL approaches. The review is focused on studies using flood detection, mapping, prediction, and monitoring; and examines methodological approaches, data fusion techniques, performance, and research gaps.

Methods: In accordance with the PRISMA 2020 guidelines, a systematic review was conducted to analyze 40 studies published from 2014-2024. Studies were selected based on pre-defined inclusion/exclusion criteria and then synthesized by ML technique, data source, application type, and terrain.

Results: Deep learning models such as U-Net and DeepResUNet had the

*Corresponding author

Email addresses: 2221128@iub.edu.bd (A. S. M Borhanul Islam), 2221047@iub.edu.bd (Sumaiya Tabassum), mdasifbinkhaled@iub.edu.bd (Md Asif Bin Khaled)

[1]These authors contributed equally to this work

highest overall performance with F1 scores up to 0.976. Tree-based ensembles such as Random Forest also proved robust across diverse terrains. Fusion of the satellites helped overcome the cloud cover, while GEE excelled in both scalability and real-time processing. This review proposes a continued research agenda with the purpose of furthering standardization of benchmark datasets, the development of models to mitigate environmental confounders, and the improved fusion of multi sensor data towards the development of reliable, scalable, and transferable flood monitoring techniques.

Keywords:
Sentinel-1, SAR, Sentinel-2, Systematic Literature Review, PRISMA 2020, Remote Sensing, Machine Learning

---

1. Introduction

Floods are considered as one of the most destructive natural disasters as it is causing widespread destruction to human lives, infrastructures, agricultural lands, and ecosystems. Floods have the potential to lead to secondary crises, such as an epidemic, by spreading contaminated water or interrupting sanitation systems. On average, floods kill more than 6,000 people and result in approximately 25.5 billion USD economic losses globally, with damage increasing by 6.3% and fatalities increasing by 1.5% per year between 1970 and 2020 (Wu et al., 2023). For instance, the Yangtze River Basin floods affected 30.2 million people and resulted in losses of 61.79 billion CNY in 2020 (Wu et al., 2023). The Ganges Delta in 2019 and Poyang Lake floods in 2020 disrupted millions of people and destroyed crops (Dong et al., 2021; Guo et al., 2021). The causes of these recurring disasters are varied, including climate change, rapid urban growth, and changes in land use. Therefore, different types of flooding such as riverine, pluvial, coastal, and flash flooding continue to threaten communities, economies, and the environment (Shahabi et al., 2020; El-Haddad et al., 2021).

To mitigate flooding effects, flood control measures include designing infrastructure (e.g., dams, levees, floodwalls) and non-infrastructure programs (e.g., forecasting, flood susceptibility maps, community preparedness). However, despite both structural and non-structural measures, traditional flood control measures still face difficulties due to the underlying reliance on experts and struggle to scale effectively across diverse regions (Nallapareddy and Balakrishnan, 2020). In this situation, satellite-based remote sensing

has become a viable alternative, as it provides high-resolution and nearly real time data for flood monitoring and prediction. The European Space Agency's Copernicus Programme, for example, introduced the Sentinel-1 and Sentinel-2 satellites to support global flood monitoring capabilities. The Sentinel-1 satellite is equipped with a C-band Synthetic Aperture Radar (SAR), which collects imagery of earth surface in both day and night, penetrating through clouds and in adverse weather conditions, using VV and VH polarizations for each of its passes. This ability makes it suitable for detecting floods, regardless of weather conditions (Jenifer and Natarajan, 2022). Sentinel-1 is also well suited to monitoring floods in dynamic environments as it has very rapid revisit times, for example, while it may only be a few days between overpasses in the floodplains of Europe or the rice regions of Asia, floods can occur and change these landscapes (Wagner et al., 2020). The Sentinel-2, by design provides vibrant mapping of water bodies and land use under cloud-free skies using multispectral bands, (Bai et al., 2021). Therefore, while optical imagery is largely impeded by cloud cover, and traditional methods of thresholding imagery contributing to error based on noise and variability seen within any region (Wu et al., 2023; Nallapareddy and Balakrishnan, 2020), using the two satellites together increases perceived reliability while monitoring the earth's surface in any condition (Lam et al., 2023; Khamphilung et al., 2023). Along with recent technological developments, machine learning (ML) and deep learning (DL) methods have improved the effective use of satellite data and allowed better flood mapping and predictions (Wedajo et al., 2024; Ghosh et al., 2024; Shahabi et al., 2020). For instance, in the Yangtze River Basin, UNet and DeepResUNet reported F1-scores of 0.976 and 0.947, respectively, which were considerably more effective than any variant of thresholding techniques (Wu et al., 2023). In addition, hybrid methods, such as swarm-optimized neural networks, have reported areas under the curve (AUC) above values of 0.93 for flood susceptibility mapping (Nguyen et al., 2023; Ngo et al., 2018). These advanced methods also resolve problems associated with SAR noise more efficiently and were able to map flooding quickly in agriculture flood assessments in Thailand, as noted (Khamphilung et al., 2023). While these models have revolutionized flood mapping, some challenges still persist. Several studies are ongoing to determine the extent of low-quality or missing datasets, the labor-intensive requirement of data preparation, and a lack of model transferability that is often devoted to single-case studies of specific areas (Ghosh et al., 2024). Additionally, many studies have focused on several flood events or localized

regions to optimize their data appropriate for secondary modelling/training for new case study applications, limiting their scalability or other applicability (Wu et al., 2023; Guo et al., 2021).

Thus, this systematic literature review seeks to integrate findings from 40 studies to address the knowledge gaps in the literature and investigate the integration of Sentinel-1 and Sentinel-2 data with ML and DL approaches. This review draws attention to the strategies identified to address data limitations, such as semi-automated generation of 5,296 strong labeled tiles (Wu et al., 2023). The review examines the strength of multi-modal data, where accuracy is improved by fusing SAR and optical imagery (Bai et al., 2021) as well as lightweight models that include real-time flooding detection (Wang et al., 2023). The main goal of this review is to introduce a set of scalable and transferable frameworks which contribute to flood monitoring including prediction and detection, thus lessening the impact of floods on human societies, economies, and ecosystems.

2. Methods

2.1. Eligibility Criteria

Table 1: Inclusion and Exclusion Criteria for Study Selection

Inclusion Criteria	**Exclusion Criteria**
1. Focused on flood monitoring and/or prediction	1. Focused only on non-flood events (e.g., drought, landslide).
2. Utilized Sentinel-1 SAR and/or Sentinel-2 multitemporal data	2. Used only non-Sentinel data sources (e.g., Landsat, MODIS).
3. Applied or discussed machine learning techniques.	3. Employed non-machine learning analytical approaches.
4. Reported quantitative performance metrics (e.g., Accuracy, F1-score, Precision) or qualitative outcomes (e.g., improved scalability, deployment success, or architectural benefits).	4. No quantitative metrics or qualitative outcomes were reported.
5. Published between 2014 (Sentinel-1 launch) and 2024.	5. Published before 2014 and after 2024.
6. English language journal articles or conference publications.	6. Non-english language journal articles or conference publications.

* Table lists inclusion and exclusion criteria for selecting studies.

Criteria for inclusion were determined to make sure the relevance at a methodological level and the applicability at a technical level. The IC/EC is

shown in table 1. The synthesis methodology categorized studies according to four primary dimensions to facilitate comprehensive analysis: (1) The ML Technique, that was the differentiation among CNN-based models, tree-based algorithms, and other ML methods; (2) The Source of Data, that was the classification of the studies as making use of Sentinel-1 exclusively, Sentinel-2 exclusively, or an integrated approach combining both sources; (3) The Type of Application, that was the discrimination between flood detection, mapping, prediction, and monitoring objectives; and (4) The Terrain/Region, that was the assessment of the performance variations across urban, rural, and mixed areas. This structured analytical framework enabled the comparison of the effectiveness of machine learning, accuracy metrics, and contextual applicability over diverse techniques, data sources, applications, and geographical settings. The synthesis approach facilitated identification of methodological strengths, performance patterns, and optimization strategies for flood monitoring and prediction utilizing Sentinel-derived remote sensing data.

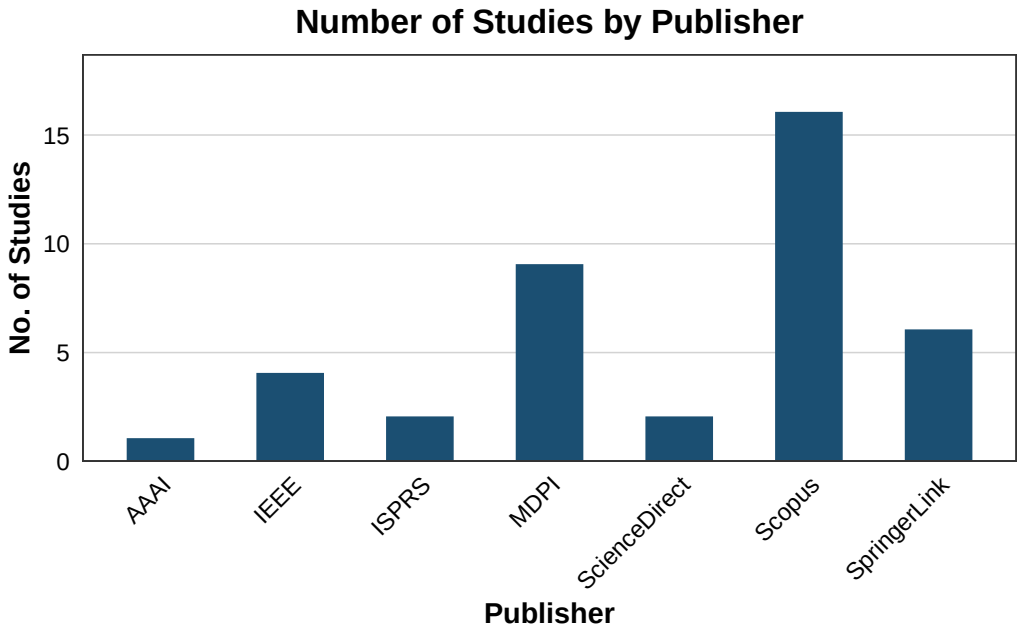


Figure 1: Number of Studies Vs Publication

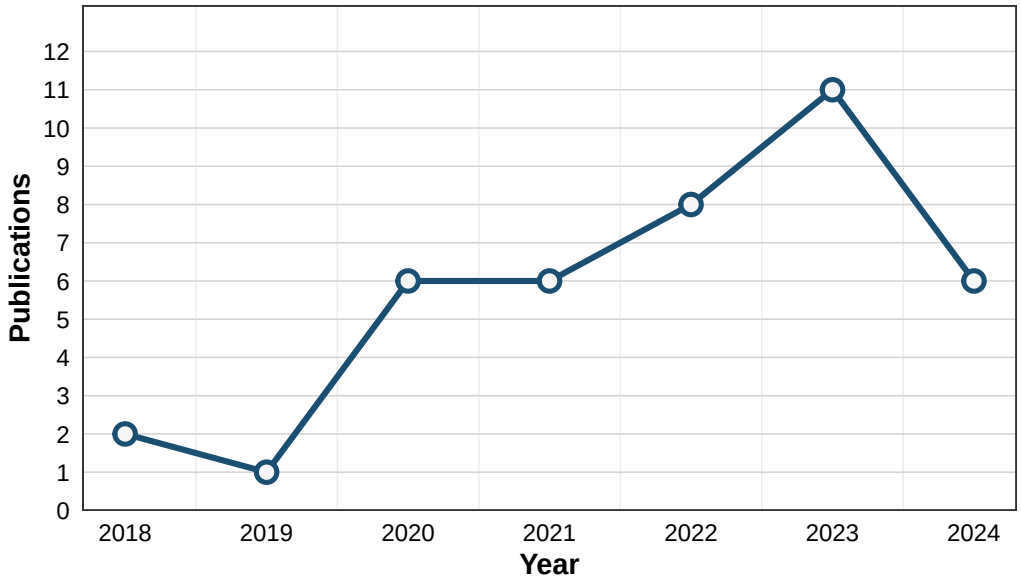


Figure 2: Number of Publications by Year on Flood Monitoring, Detection and Prediction Studies (2018–2024)

2.2. Information Sources

The search for studies relevant to this systematic review took place in the following databases: IEEE Xplore, ScienceDirect, SpringerLink, Scopus, MDPI, Taylor & Francis Online, and Wiley Online Library. We used a predetermined search string and filtered the range of duration from 2014 to 2024, corresponding to the launch of Sentinel-1 and Sentinel-2 satellites. A table is provided with search strings used for each database (see Table 2).

2.3. Search Strategy

The search strategy for the identification of studies from the selected databases was prepared in a Google Sheets file. The study type, year of publication, language of preference, search string, and the name of the database were determined, and the searching was carried out as per this plan. The search strings used in all these databases, as presented in Table 2 were modified to the respective syntax rules of the different databases while the logical formulation was maintained intact.

Table 2: Database search strategy for flood monitoring studies using remote sensing and machine learning.

Databases	**Search String**
IEEE Xplore	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
ScienceDirect	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
SpringerLink	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
Scopus	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
MDPI	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
Taylor & Francis	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")
Wiley Online Library	("flood monitoring" OR "flood detection" OR "flood prediction") AND ("remote sensing" OR Sentinel OR SAR OR "satellite imagery") AND ("machine learning" OR "deep learning" OR "neural network")

* Search conducted from 2014 to 2024 in English.

2.4. Selection Process

The selection strategy followed the process discussed previously to make sure it met the inclusion and exclusion criteria. Two reviewers, reviewer A and reviewer B, independently screened the titles and abstracts of the identified records using that criteria. Both reviewers then screened the full articles by using the inclusion and exclusion criteria and tried to resolve differences by discussion with a senior reviewer, reviewer S, or by concurrence. To avoid differences in the screening process, a new verification step was incorporated, using a large language model (LLM). The articles accepted into the review were run through the model after human screening was completed. The model was provided with a deterministic structured prompt which had the same inclusion and exclusion criteria (e.g., "must use Sentinel-1 and/or Sentinel-2," "must use machine learning" , "must analyze flood events," etc.) and were instructed to identify any studies which did not meet these criteria based on the title and abstract. It should be made clear that the LLM was not

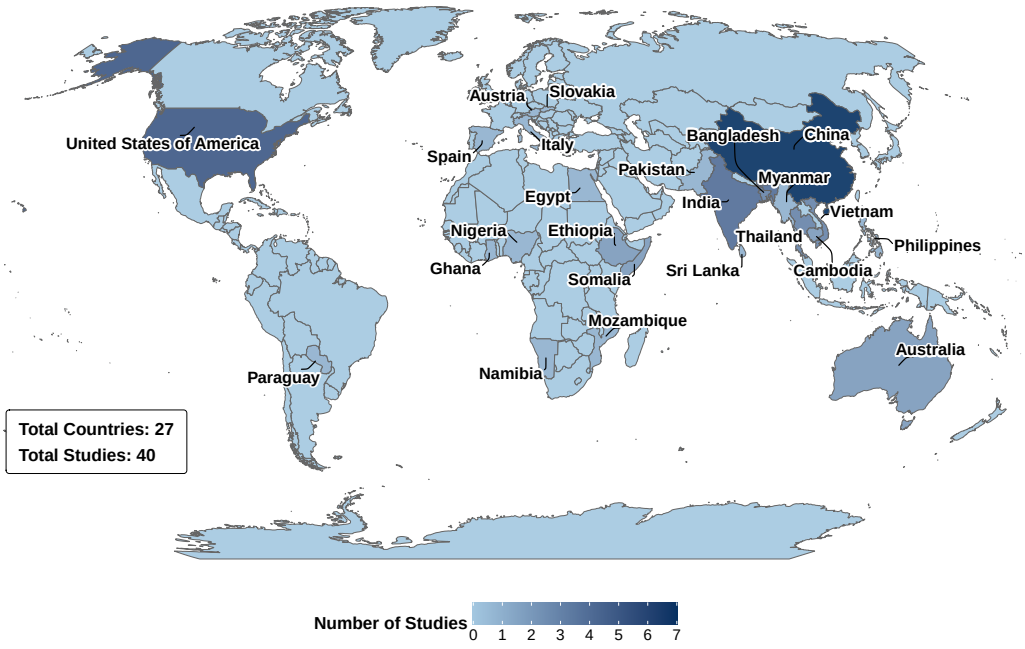


Figure 3: Geographical distribution of included studies. Choropleth gradient by number of studies per country (N=40; excludes 5 global studies).

intended to make decisions for the reviewers, but to ensure the consistency of their decision-making and highlights where possible chances of human error. All discrepancies identified by the LLM were decided upon by reviewer S who made the final decision. In this way the pattern-matching capabilities of the LLM were used and kept the process methodologically controlled.

2.5. Data Collection Process

Data was extracted manually by reviewer A and reviewer B. A data extraction file was developed prior to the process based on the disintegrated segments of 10 research questions, which are stated in Table 3.

Table 3: Research questions and corresponding key aspects for flood monitoring and prediction studies using Sentinel data and machine learning.

Research Questions	**Key Aspects**
RQ1: Effectiveness of Sentinel-1 SAR for flood detection under environmental and atmospheric conditions, including challenges in urban and vegetated terrains?	1) Sentinel-1 SAR in flood mapping; 2) Environmental and atmospheric; impacts; 3) SAR-based detection techniques; 4) Performance metrics; 5) Urban and vegetated terrain; challenges
RQ2: Integration of Sentinel-1 SAR and Sentinel-2 optical data for enhanced flood mapping accuracy, including trade-offs of single- vs. multi-sensor approaches?	1) SAR and optical integration techniques; 2) Single- vs. multi-sensor comparison; 3) Trade-offs in flood mapping; 4) Data integration challenges; 5. Best practices for accuracy
RQ3: Comparison of CNNs and advanced ML techniques vs. traditional rule-based and statistical methods for flood detection and prediction using Sentinel data?	1) CNNs and ML models; 2) Traditional rule-based/statistical methods; 3) Performance metrics; 4) CNNs vs. traditional methods; 5) Method selection recommendations
RQ4: What are the impacts of different CNN architectures (e.g., U-Net, DeepLabV3+) and training strategies (e.g., data augmentation, loss functions) on the performance of flood detection models using Sentinel-1 and Sentinel-2 data?	1) CNN architectures; 2) Training strategies; 3) Overfitting challenges; 4) Dataset and training details; 5) Performance metrics; 6) Architecture comparison; 7) Optimization recommendations
RQ5: How well do machine learning-based flood detection models generalize to unseen flood events and across different geographic regions, and what factors influence their transferability?	1) Generalization testing; 2) Datasets for unseen events/regions; 3) Performance on diverse regions; 4) Transferability factors; 5) Transferability techniques; 6) Generalization recommendations
RQ6: What are the operational implications of adopting machine learning-based models for real-time flood monitoring, and how scalable and computationally efficient are these models for large-scale applications?	1) ML for real-time monitoring; 2) Scalability and efficiency; 3) User interface design; 4) Operational challenges; 5) Scalability optimizations; 6) Cost analysis; 7) Real-time recommendations

Table 3 –

Research Questions	Key Aspects
RQ7: How accurate and scalable are fully automated processing chains like the Sentinel-1 Flood Processor (S-1FP) for flood detection, and how do they compare to machine learning-based approaches?	1) Automated processing chains; 2) Accuracy and scalability metrics; 3) Comparison with ML approaches; 4) Advantages and disadvantages; 5) Integration with ML; 6) Recommendations for automation
RQ8: How do dual polarization modes (VV and VH) in Sentinel-1 data influence the accuracy of flood detection models, and what are the trade-offs in using different polarization combinations?	1) VV and VH polarization modes; 2) Impact on accuracy; 3) Polarization trade-offs; 4) Polarization challenges; 5) Optimization recommendations
RQ9: How does the integration of auxiliary data such as Digital Elevation Models (DEMs) and the Height Above Nearest Drainage (HAND) index improve the accuracy of flood detection using Sentinel data?	1) Auxiliary data used; 2) Integration methods; 3) Accuracy impact; 4) Integration challenges; 5) Accuracy benefits; 6) Recommendations for auxiliary data
RQ10: What role do cloud-based platforms like Google Earth Engine (GEE) play in enabling large-scale, automated flood detection, and how can they be optimized for processing multi-sensor data in real time?	1) Cloud platforms' role; 2) Automation benefits; 3) Real-time processing optimization; 4) Platform challenges; 5) Large-scale applications; 6) Real-time recommendations

2.6. Data Items

Primary and secondary outcomes are listed in Table 4, which was determined based on our research questions.

2.6.1. Other Variables

Other than the research questions, we collected some additional variables from the studies for synthesis purposes. These variables were: Authors, Year Published, Title, Publisher, Country/Region of study, Study Design, Sample Size, data characteristics, study area (urban or rural), flood type, outcome, validation method, negative findings, and funding source.

2.7. Study Risk of Bias Assessment

We developed a structured process to determine the risk of bias in the studies which we had reviewed systematically. As there are no pre-existing risk-of-bias tools specifically for machine learning (ML) or remote sensing studies, we decided to utilize the ROBINS-E tool (Xu et al., 2023) also faced a similar situation of not having an established risk of bias tool for an emerging technological discipline and developed their own assessment tool. ROBINS-E was originally developed for non-randomized studies of exposures, it

Table 4: Primary and Secondary Outcomes.

Outcome	Description
Primary Outcome	The effectiveness of machine learning models in detecting floods using Sentinel-1 and Sentinel-2 data, compared to traditional rule-based and statistical methods, including an analysis of their performance metrics and their ability to handle different environmental and terrain complexities.
Secondary Outcome 1	The influence of different ML model architectures and data integration strategies on the accuracy and robustness of flood detection models, including an analysis of the trade-offs and best practices for combining Sentinel-1 and Sentinel-2 data.
Secondary Outcome 2	The transferability and generalization of ML-based flood detection models to unseen flood events and across diverse geographic regions, and the factors that influence their performance in new contexts.
Secondary Outcome 3	The operational implications, scalability, and computational efficiency of using ML models for real-time flood monitoring, including an examination of how cloud-based platforms facilitate large-scale applications.
Secondary Outcome 4	The impact of different data characteristics, such as Sentinel-1 polarization modes and the integration of auxiliary data, on the accuracy of ML-based flood detection models.

was chosen instead of other tools like QUADAS-2 because ML flood monitoring studies are observational, where the "exposure" refers to using a specific ML technique or data combination for prediction or monitoring. From the seven ROBINS-E domains, we chose five domains which were most relevant to our study, including confounding (D1), selection of data (D2), missing data (D5), measurement of outcome (D6), and selection of reported results (D7). Each domain was rated using signaling questions specifically designed for ML and remote sensing applications and evaluated for low risk of bias (0 = No, 1 = Probably No, 2 = Yes) and scored per domain, and compressed into a score of Low risk (i.e., ≥ 6), Some Concerns (i.e., 4–5), and High risk of bias (< 4). The overall risk of bias for each study was determined by the highest risk in any critical domain. This meant that even one serious flaw (i.e., high risk in confounding) increased the overall risk. The risk of bias for all included studies was assessed separately and independently by reviewer A and reviewer B. Before using our adapted ROBINS-E tool, we first completed a calibration exercise on a sample of 5 studies. After resolving discrepancies with reviewer S, we identified consistent and uniform scoring. Inter-rater reliability using Cohen's Kappa was calculated for the sample of studies yielding a coefficient of $\sim$0.82, suggesting that there was substantial agreement between reviewers. Following the independent reviews of all 40 studies, inter-rater reliability was again calculated using Cohen's Kappa to be $\sim$0.84, suggesting substantial agreement. Discrepancies for all reviews were resolved through discussion and final determination by the supervising instructor for the consensus score reported in this paper.

- **Domain 1: Bias due to Confounding (D1)**: This domain focused on whether studies dealt with the effective control of confounding factors that could influence the performance of flood detection/prediction methods. Confounders included the various preprocessing steps (for example, image calibration, speckle filtering), environmental (for instance, terrain complexity, vegetation density, rainfall intensity), and differences in validation datasets. Signaling questions examined whether studies accurately and clearly established what they were comparing, the design of the study accounted for the quality of data, and/or whether they addressed the possible role of environmental variability.
- **Domain 2: Bias in Selection of Data (D2)**: This domain investigated the appropriateness of the data selection process, including selection of study areas, flood events, and datasets (e.g., Sen1Floods11 vs local datasets). Signaling questions considered whether the datasets were representative of the flood monitoring context and if selection biases (e.g., restricting to specific flood types and/or regions) had been avoided.
- **Domain 5: Bias due to Missing Data (D5)**: This domain focused on handling missing satellite data, such as missing information from cloud cover for Sentinel-2 or gaps in Sentinel-1 during certain time periods. It also addressed the problems of incomplete reporting for preprocessing steps. Signaling questions examined whether the studies used imputation, or strategies to mitigate missing data, and whether these were reported alongside their impact on performance.
- **Domain 6: Bias in Measurement of Outcome (D6)**: evaluated at how valid ground truth data were, and whether each performance metric (e.g., F1-score, IoU) were relevant to the outcome of interest. Signaling questions examined whether ground truth was based on reliable reference data (e.g, proper in-situ measurements, high-resolution imagery) and whether the reporting of the metrics was consistently applied across the various methods being compared.
- **Domain 7: Bias in Selection of Reported Results (D7)**: addressed whether studies reported outcomes that were relevant or only from the models/metrics that performed the best. Signalling questions considered whether reports were comprehensive, including error rates, secondary outcomes and limitations.

The adapted ROBINS-E tool also used a quantitative scoring framework to provide uniform evaluation, using signaling questions, and risk thresholds tailored to the specific requirements of ML-based flood monitoring and prediction studies.

2.8. Effect Measures

Following the PRISMA 2020 guidelines, we chose the effect measures based on their prevalence and importance in studies that assess machine learning algorithms for flood monitoring and prediction using satellite imagery from Sentinel-1 and Sentinel-2. The selected effect measures for this review are Accuracy, Precision, Recall, and F1 score. These were the most reported assessment measures in the reviewed literature. Accuracy is the number of observations correctly identified, including both floods and non-floods from

the total instances evaluated. It is a simple way to measure overall model performance and is helpful for evaluating models during classification tasks in flood detection (Powers, 2020). Precision means that the predicted floods matched to the extent of how many they were really there. This step is crucial for flood monitoring as it can help the system save resources and avoid alerts by no longer creating false alarms (Powers, 2020). Recall or sensitivity indicates the number of real flood events that the model was able to detect correctly. This is of utmost importance when the data contains a lot more non-flood cases compared to flood ones. It is necessary to make sure that most flood events are detected according to (Powers, 2020). . Recall or sensitivity shows how many actual flood events are correctly the model was able to detect. It's especially important when the data has many more non-flood cases than flood ones. It is important to ensure that most flood events are detected, according to (Powers, 2020). The F1 score is the harmonic mean of Precision and Recall. Its primary purpose is to help balance the trade- offs between these two assessment measures. These metrics are especially valuable in flood monitoring studies with imbalanced datasets since they provide a single metric to evaluate model performance, considering both false positives and false negatives equally (Zuhairi et al., 2024). These metrics convey the classification performance of models trained with machine learning on Sentinel-1 and Sentinel-2 data for flood monitoring and prediction tasks.

2.9. Synthesis Methods

2.9.1. Eligibility for Synthesis

The eligibility process for the synthesis was designed based on four major analytic dimensions: the classification of machine learning approaches, data source, application type and terrain characteristics. All studies needed to satisfy the inclusion criteria, which were eligibility for syntheses as: (1) used the Sentinel-1 and Sentinel-2 data source as the primary source, (2) used any of the machine learning methods developed to address flood-related applications, (3) provided a measurable performance metric, and (4) described the application and terrain characteristics in a reasonable manner. Relevant syntheses did not include studies with insufficient method detail and / or comparable performance metrics. We did use these qualitative studies to create other analytic groups. This eligibility approach to synthesis allowed for a descriptive foundation as well as comparative analysis of machine learning methods, while holding comparative analytic consistency through the eligibility characteristics.

2.9.2. Data Preparation

This section addresses the issue of inconsistent reporting, differences in reporting metrics, and impacts of study formatting. Any summary statistics that were not directly reported were calculated as a metric when possible (e.g., F1 scores were calculated from precision and recall), or were eliminated from quantitative synthesis where necessary. Data was standardized by converting all performance metrics into consistent decimal formats (specifically, 0-1 scale for accuracy and F1 score). For studies where multiple models were reported, the models with best scores were used for the primary synthesis with the option to return to include all the data for secondary analyses. The data allowed for both quantitative comparisons, which included tabular summaries of performance metric

scores reported by group, and qualitative thematic analysis of methodological insights. Consequently, a comprehensive assessment of the effectiveness of machine learning was guaranteed through the various techniques, data sources, applications, and terrain contexts.

2.9.3. Tabulation and Visualization

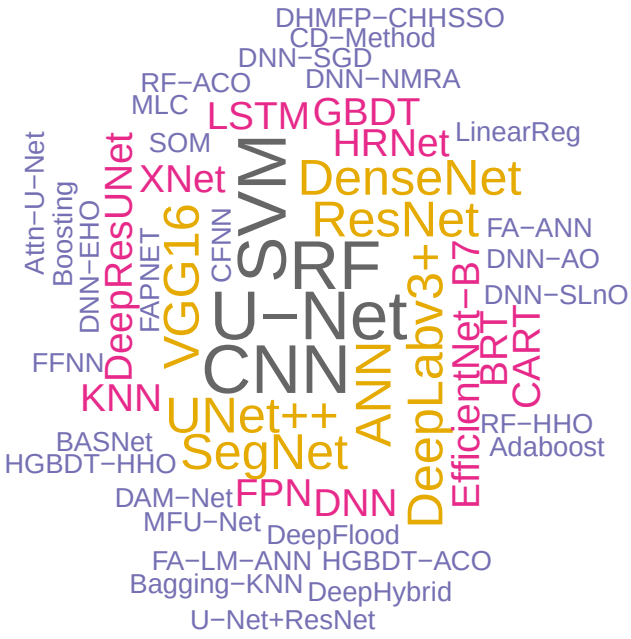
We created both tabular and graphical formats of the results to enable a better comparative analysis and employed two distinct formats for tables: (1) A summarized table of individual study characteristics that records crucial metadata (check table 5)) and performance metrics (see table 7)); and (2) A table that merges performance metrics and aggregates statistical measures among synthesis groups (refer to table 8)). Along with the tables, we plotted some graphs. We presented the distribution of studies by terrain type (Rural, Urban, Both) and by data source category (Sentinel-1, Sentinel-2, or Combined) using pie charts. Grouped bar plots helped visualize the highest performance metrics across studies by task (Accuracy, F1, Precision, and Recall). We also used radar plots for the GRADE Certainty Assessment across five study outcomes (PO1, SO1-SO4) in six domains. This structured way of presenting information helped us evaluate how effective machine learning is across various techniques, data sources, applications, and geographic areas while keeping our analysis clear and understandable.

2.9.4. Synthesis Methods Used

Meta-analysis was not appropriate due to substantial heterogeneity across the 40 studies. Variability in the study designs, variable performance measures, and no standardization of effect measures contributed to the decision. Therefore, a combination of narrative synthesis and descriptive statistical approaches was employed within the context of preestablished synthesis groups. The narrative synthesis engaged in making inferences on trends in quantitative performance and qualitative characteristics across studies. Performance measures were summarized by the synthesis group to identify characteristics and trends; for example, CNN models had a higher average F1-scores (≈ 0.92) compared to tree-based models (≈ 0.91). Thematic analysis was performed on the qualitative findings. Descriptive statistical synthesis provided organized quantification of performance measures across groups via measures of central tendency (e.g., mean, median) and dispersion (range, and standard deviation). All measures were standardized to a 0–1 scale, where measures were calculated as needed (e.g., F1 scores calculated from precision and recall for Lam et al. (2023); De La Cruz et al. (2020); Tsyganskaya et al. (2018); Ngo et al. (2018); Guo et al. (2021)). Studies without quantitative performance measures were omitted from statistical summaries yet were included in qualitative synthesis. The analyses were undertaken using Python and the pandas library.

2.9.5. Investigation of Heterogeneity

We were unable to apply standard statistical methods such as meta-regression or subgroup analyses due to the considerable methodological heterogeneity present across machine learning architectures, combinations of data sources, types of application tasks, and performance metrics. In order to evaluate potential causes of heterogeneity we will compare performance metrics amongst specified synthesis groups: type of machine learning


DHMFP−CHHSSO
CD−Method
DNN−SGD
RF−ACO
DNN−NMRA
MLC
LSTM
GBDT
LinearReg
SOM
HRNet
Attn−U−Net
UNet
XNet
DenseNet
Boosting
DNN−EHO
ResNet
FA−ANN
SVM
DeepResUNet
VGG16
CFNN
RF
DNN−AO
FAPNET
DeepLabv3+
EfficientNet−B7
DNN−SLnO
U−Net
BRT
CART
FFNN
CNN
ANN
KNN
UNet++
RF−HHO
BASNet
SegNet
Adaboost
HGBDT−HHO
DAM−Net
FPN
DNN
MFU−Net
DeepFlood
FA−LM−ANN
HGBDT−ACO
Bagging−KNN
DeepHybrid
U−Net+ResNet

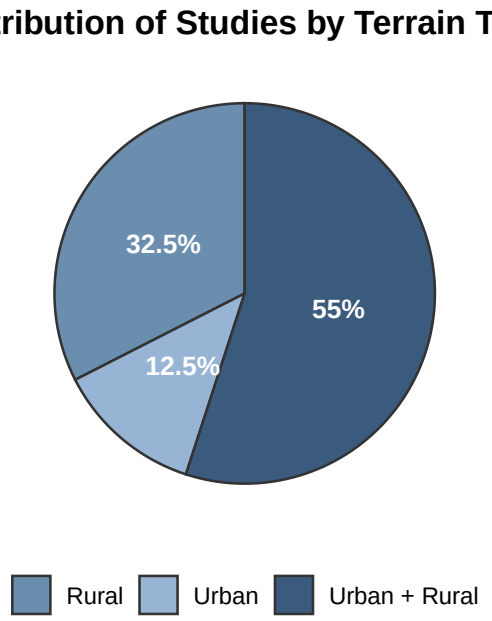


Figure 4: Geographic Scope of Flood Monitoring and Prediction Research. This chart demonstrates that most studies reviewed adopt a broad scope, with over half the research covering Both urban and rural terrain types.

(CNN-based, tree-based, other); data source used (Sentinel-1, Sentinel-2, composite); type of application (detection, mapping, prediction and monitoring); and type of terrain (urban, rural, mixed) and look for differences in average performance metrics (accuracy and F1-score). Our goal is to find the patterns that inform potential sources of the heterogeneity in studies. Determinants of heterogeneity such as task complexity, model complexity, quality of data and type of terrain will be highlighted in visualizations such as bar charts, boxplots, scatter plots and heatmaps.

2.9.6. Sensitivity Analyses

The robustness of synthesized findings will be examined through three pre-specified sensitivity analyses. First, qualitative themes will be re-evaluated after excluding qualitative-only studies to evaluate if there have been any changes in primary conclusions. Second, descriptive statistical summaries will be recalculated after excluding studies that may have overly influenced results as statistical outliers (e.g., from interquartile ranges of F1-scores) and explore the effect on the average performance metric and group rankings. Lastly, where possible, primary conclusions about F1-scores will be compared to those which derived from a different metric (e.g., IoU for segmentation tasks), so that any variations in performance trends across synthesis groups can be assessed.

Non-Quantitative Studies Analysis: exclusion of three studies with no quantitative metrics (Zhao et al., 2023; Tiampo et al., 2021; Wagner et al., 2020) from qualitative themes depicted a minimal effect on major findings and qualitative insights from these studies aligned with the broader dataset.

2.10. Reporting Bias Assessment

To assess the biases regarding missing results (due to selective reporting) in the quantitative synthesis of the 40 included studies, the Domain 7 (selection of reported results) of the modified ROBINS-E tool was used, as described in section 2.7. The Domain was considered to evaluate whether studies did report all relevant outcomes or, selectively, it reports only best-performing models/metrics according to the signaling questions, rated with one of three values: 0 (No), 1 (Probably No), and 2 (Yes). Therefore, the systematic literature review is a quantitative review without meta-analyses,the usual statistical tests to infer publication bias would not be applicable: for example, funnel plots or related tests for asymmetry were not used. Assessment of reporting bias is therefore adequately assessed with the broad evaluation of Domain 7 scores together with a narrative synthesis, and ensuring that selective reporting does not excessively affect synthesis outcomes.

2.11. Certainty Assessment

The GRADE (Grading of Recommendations Assessment, Development and Evaluation) system was used to evaluate certainty of evidence originating from the systematic literature review. This framework evaluates domains such as risk of bias, inconsistency, indirectness, imprecision, and publication bias, as well as other considerations that would warrant a higher rating (i.e., large effects or dose-response gradients). Domains were rated based on predefined criteria: for risk of bias, we rated 1 (low risk) if $\geq$50% of studies had clear and robust approaches (e.g., per ROBINS-E equivalents), rated 0 (moderate risk) for mixed ratings, and rated -1 (serious/critical risk) if predominant concerns existed. For inconsistency, we rated 1 (consistent) if $\geq$50% of studies had stable results, rated 0 for moderate variation and -1 for wide or opposing variations in effect direction. Indirectness received 1 (direct) if $\geq$50% of studies had similar populations, interventions, or comparators, as defined by the Population, Intervention, Comparator and Outcome (PICO) framework, relative to Sentinel-1/2 based ML flood detection; rated 0 if partially relevant; and rated -1 if indirectly relevant. For imprecision, we rated 1 (precise) if $\geq$50% of studies had large datasets (>200 images/tiles) or small confidence intervals, rated 0 for moderate impreciseness, and rated -1 (imprecise) if sample sizes were small (<200 images) or had unclear intervals.

3. Results

3.1. Study Selections

3.1.1. Screening Results

Following the PRISMA 2020 guideline, we collected 2,064 records through database sources. We merged all the records in a single sheet and deduplicated 219 records, We

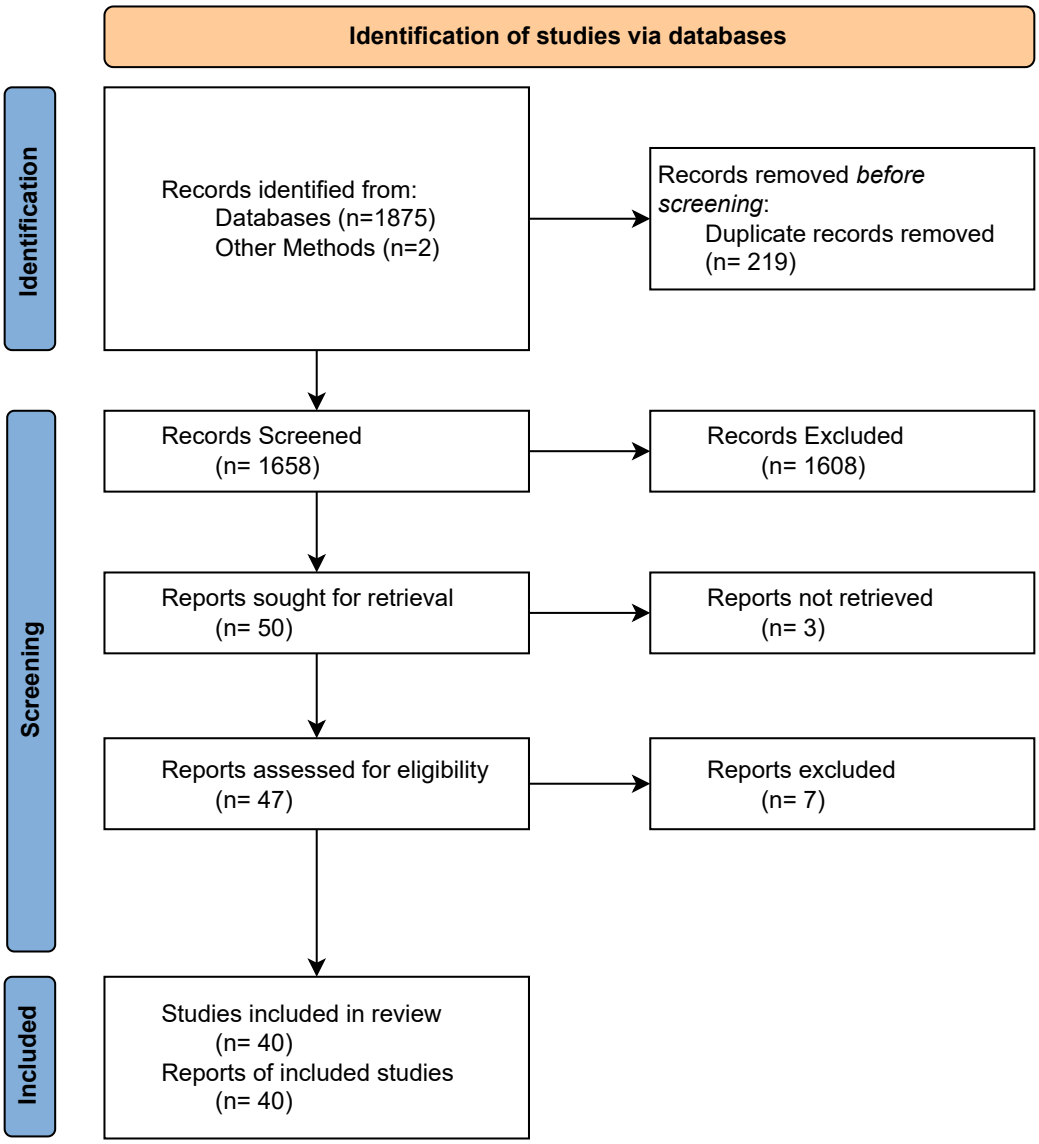


Figure 5: PRISMA 2020 flow diagram.

screened the title and abstracts of 1,658 records resulting in 50 records that were assessed for eligibility at the full-text level. Of the 50 records, three records that we assessed were non-accessible. The remaining 47 studies met the inclusion criteria. After full text assessment, 7 reports were excluded due to reasons mentioned in section 3.1.2 and 40 studies were included in the qualitative and quantitative synthesis, as shown in Figure 5

3.1.2. Excluded Studies

Following the screening process, 47 records were sought for retrieval. After assessing the full text of all reports, 7 of them were excluded. This final selection process resulted in 40 unique studies included in the review, all of which were sourced from 40 reports.

3.2. Study Characteristics

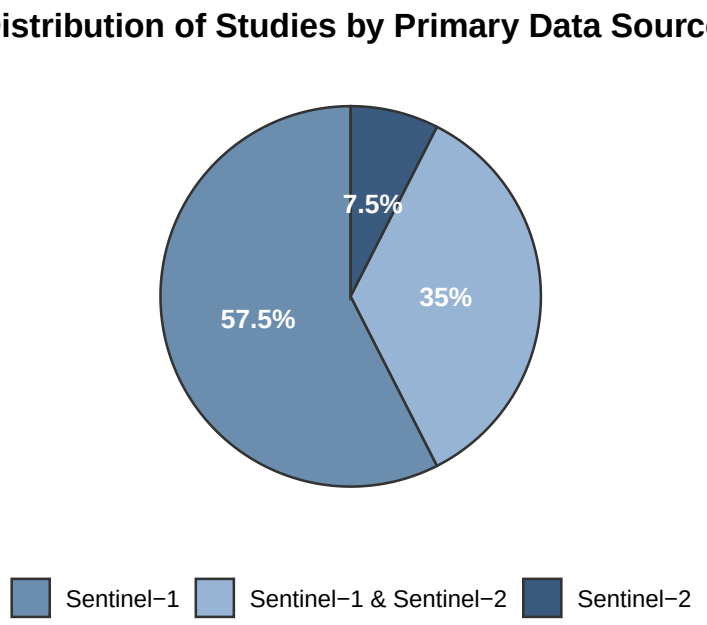


Figure 6: Distribution of Studies by Data Source.

The key strategies employed by the 40 studies are specified in Tables 5 which present details of study design, study setting and machine learning techniques. Most studies (47.5% of studies) were published between 2022-23, indicating a recent thrust in operational, The 40 studies span at least 34 countries, with the highest concentration in Asia and Europe (See Figure 3). Figure 6 shows the breakdown of studies included by data source with Sentinel-1 (SAR) data (57.5%), Sentinel-2 (multispectral, 7.5%) data, and combined sources (35%).

Table 5: Study characteristics with ML techniques.

Sl.	Study	Study Design	Setting	ML Techniques
1	Wu et al. (2023)	CNN-based flood detection and mapping	Rural	FCN-8, SegNet, UNet, DeepResUNet

Table 5 –

Sl.	Study	Study Design	Setting	ML Techniques
2	Wedajo et al. (2024)	Comparative ML models for flood prediction	Rural (semi-arid)	Random Forest, Linear Regression, LSTM, SVM
3	Dong et al. (2021)	CNN vs. traditional methods for flood monitoring	Urban, rural	HRNet, DenseNet121, SegNet, ResNet101, DeepLabv3+, BTS, Otsu
4	Lam et al. (2023)	ML algorithms for flood mapping	Rural	CNN, MLP, RF
5	Stateczny et al. (2023)	Deep hybrid model using satellite imagery	Urban	DHMFP–CHHSSO, DBN, RNN, LSTM, SVM, Bi-GRU, FCNN
6	Tanim et al. (2022)	Supervised / unsupervised ML for flood detection	Urban	RF, SVM, MLC, Change Detection
7	Fraccaro et al. (2022)	AI + Sentinel-1 + geospatial data for flood detection	Urban, rural	ResNet-based UNet
8	Ghosh et al. (2024)	Nested UNet on Sentinel-1 SAR data	Urban, rural	UNet++, EfficientNet-B7
9	Bhadra et al. (2020)	CNN trained on Sentinel-1/2 for flood detection	Rural	CNN
10	Patil et al. (2023)	CNN models on Sentinel-2 for flood detection	Urban, rural	VGG16, DenseNet, GoogleNet
11	Zhao et al. (2023)	Multi-temporal SAR case study	Urban, rural	Multi-Attention-UNet
12	Nallapareddy and Balakrishnan (2020)	Sentinel-1 SAR + ML for flood analysis	Urban, rural	Feed-Forward NN, Cascade NN
13	De La Cruz et al. (2020)	U-Net on multi-temporal Sentinel-1A SAR	Urban, rural	Modified U-Net
14	Bai et al. (2021)	Deep learning models on benchmark datasets	Mixed	BASNet, Deeplabv3+, FCN-ResNet50, U2Net
15	Nemni et al. (2020)	CNN vs. classical ML for flood mapping	Urban, rural	U-Net, XNet, ResNet backbone
16	Ghosh et al. (2022)	UNet + FPN architectures for flood detection	Urban, rural	UNet, FPN, EfficientNet-B7
17	Shahabi et al. (2020)	Sentinel-1 + ML for susceptibility mapping	Rural	KNN variants, Bagging Tree
18	Zhang et al. (2023)	MFU-Net for water segmentation (Sentinel-1/2)	Mixed	MFU-Net, SAR-Net, MSI-Net, WI-Net, SWI-Net, BASNet, DUPNet
19	Nguyen et al. (2023)	Hybrid DL + swarm optimization for susceptibility	Rural	DNN variants (NMRA, SLnO, EHO, AO, SGD)
20	Ebadati et al. (2024)	Comparative remote sensing methods	Urban	RF, HGBDT, HHO, ACO, VGG-16

Table 5 –

Sl.	Study	Study Design	Setting	ML Techniques
21	Islam et al. (2022)	FAPNET model with SAR fusion and patch augmentation	Urban, rural	FAPNET, UNET++, ATTUNET, LINKNET
22	Jenifer and Natarajan (2022)	Dual patch-based FCN with feature fusion	Mixed	DeepFlood FCN, RF, NDWI, ResNet-50
23	Sghaier et al. (2019)	Multimodal optical/radar flood monitoring	Urban, rural	SFS, Modified Water Index, AF, SOM
24	Tiampo et al. (2021)	Sentinel-1 SAR + ML for depth/extent detection	Urban	Thresholding, Deep CNN
25	El-Haddad et al. (2021)	Comparative ML techniques for susceptibility mapping	Rural	BRT, FDA, GLM, MDA
26	Khan et al. (2024)	ML + GEE for real-time flood impact	Rural	ANN, RF, SVM
27	Tsyganskaya et al. (2018)	Time series classification (pixel + object-based)	Rural	Random Forest
28	Wagner et al. (2020)	Sentinel-1/2 flood event analysis	Urban, rural	Thresholding, CD, ML, Time Series
29	Prakash et al. (2024)	ML + AHP-MCE for flood monitoring	Rural	CART, RF, SVM
30	Ngo et al. (2018)	GIS + ML for flash-flood modeling	Mixed	FA-LM-ANN, SVM, CT
31	Yu et al. (2023a)	ML + Sentinel-1A flood inventory framework	Mixed (urban focus)	RF, GBDT, KNN, SVM, ANN
32	Tiampo et al. (2022)	Thresholding + DeepLabv3+ on Sentinel-1A/B	Urban	Thresholding, DeepLabv3+
33	Saleh et al. (2024)	CNN-based change detection for flood mapping	Urban, rural	DAM-Net, SiamNet variants, SwinSUNet
34	Guo et al. (2021)	Flood monitoring using OBIA and Decision Tree for water extraction	Urban, rural	Decision Tree, (OBIA), Fuzzy Classification Method, CA-Markov Model
35	Kim et al. (2021)	DL model for water body extraction (geospatial)	Urban, rural	Customized U-Net
36	Khamphilung et al. (2023)	LULC classification pre/post 2019 flood	Rural	RF, KDTree KNN, MLC
37	Thapa et al. (2022)	NDVI/NDWI + CART + mobile image verification	Rural	CART, SVM, RF
38	Yu et al. (2023b)	FWSARNet flood detection model	Urban, rural	FWSARNet, ViT, ResNet, Segformer
39	Wang et al. (2023)	MSH-SITS + lightweight U-Net for flood detection	Rural	LSFUnet, Enet, Linknet, U2Net
40	Wu et al. (2022)	Multi-Scale Deeplab on dual-pol SAR	Urban, rural	MS-Deeplab (MobileNetV2 + DeepLabv3+)

3.3. Risk of Bias in Studies

The risk of bias (RoB) in all 40 studies was evaluated with the adapted ROBINS-E tool across five domains: confounding (D1); selection of data (D2); missing data (D5); measurement of outcome (D6); and selection of reported results (D7). Each domain was rated using the specific signaling questions, with scores ranging from 0 (high risk) to 2 (low risk). The overall RoB was determined by the highest risk designation across all domains, with particular emphasis on critical domains (D1, D2, and D6). For the certainty assessment (Section 3.7), overall study-level RoB ratings (11/40 low risk, 29/40 some concerns, 0/40 high risk) are used, while domain-specific ratings provide supplementary methodological insights.

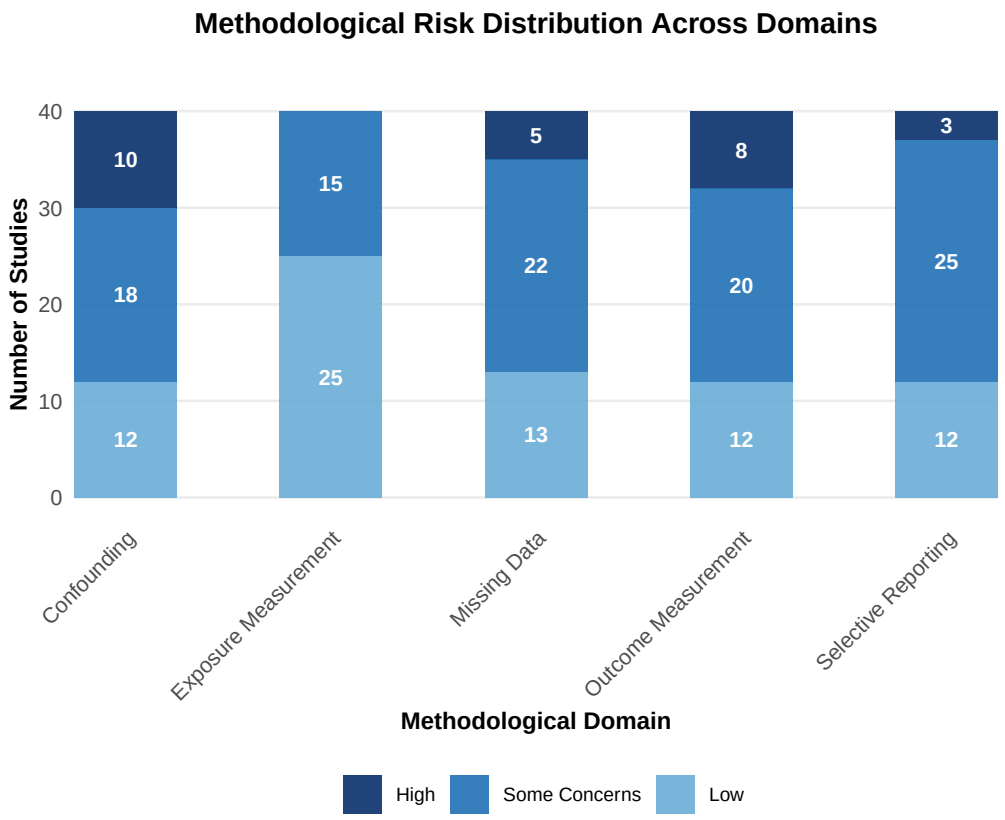


Figure 7: Grouped percentage chart showing the proportion of studies with low risk, some concerns, and high risk across different bias domains.

For the selected 40 studies, 11 (27.5%) were classified as low risk, 29 (72.5%) were rated as some concern and 0 (0%) were rated as high risk. The risk levels varied across different domains. For domain 1, 18 studies (45%) were rated as high risk due to factors not controlling for environmental confounders, such as environmental rainfall patterns or characteristics of terrain or changing land use. Domain 2, had 0 (0%) rated high risk with concerns in 8 studies (20%) and mostly due to concerns for not clearly establishing characterization of satellite data quality or minor potential for misclassifications. Overall, missing data (D5) was not an issue as only 5 studies (12.5%) were rated high risk suggesting that most studies addressed the issue of missing data or relied on fully completed data sets. 3 studies (7.5%) were classified as high

risk for outcome measurement (D6) largely due to differences in measurement. Selection of significant results (D7) had only 1 study (2.5%) as high risk, which indicates overall level of transparency.

Table 6: Justifications for risk of bias judgments for Sentinel-based flood detection studies.

ID	Study	Domain	Risk Level	Justification
1	Wu et al. (2023)	D1	Some Concerns	Partial control for confounders like rainfall; unclear reliability in terrain or land use measurements; no evidence of serious uncontrolled confounding.
3	Dong et al. (2021)	D1	High	No control for key confounders such as rainfall or terrain; unreliable or absent confounder measurements; evidence of serious uncontrolled confounding.
4	Lam et al. (2023)	D5	Low	No or minimal missing data; gaps unrelated to outcomes; predictors of missingness fully addressed (e.g., via multi-source data fusion).
9	Bhadra et al. (2020)	D6	Some Concerns	Unclear consistency in outcome measurement; potential assessor awareness of exposure; possible but unclear bias from exposure knowledge.
15	Nemni et al. (2020)	D2	Low	Strong characterization of Sentinel-1 SAR data for flood events; no or minimal risk of exposure misclassification (e.g., effective cloud-penetrating SAR).
20	Ebadati et al. (2024)	D7	Low	Pre-specified analytical plan evident; no selection of exposure, outcome, or estimates based on results; comprehensive reporting.
25	El-Haddad et al. (2021)	D7	High	No pre-specified plan; potential selection of exposure measures, outcomes, or estimates based on favorable results.
33	Saleh et al. (2024)	D1	High	No control for confounders; unreliable confounder measurements; clear evidence of serious uncontrolled confounding.

The risk of bias assessment revealed that confounding control and exposure measurement remain the primary methodological challenges in Sentinel-based flood detection studies. While the majority of studies demonstrated acceptable methodological quality with "some concerns," the significant proportion of high-risk assessments in critical domains highlights the need for improved standardization in confounder control and satellite data validation protocols.

3.4. Results of Individual Studies

Individual study results were summarized using the performance metrics: Accuracy, Recall, F1-score and Precision, for a total of 40 studies, as shown in a summary table 7. This summarizing table includes study-specific metrics, models (e.g. U-Net, Random Forest, DeepResUNet), data sources (Sentinel-1, Sentinel-2, combined), performance metric and scores, and risk of bias rating (high risk, some concerns, low risk). Due to heterogeneity in models, types of flooding and validation methods averaging performance metrics across studies is not feasible or meaningful at this stage so they are presented individually to feed into the synthesis in Figure 5

Table 7: Performance Metrics of Individual Studies

SL	Study	Data Source	Best Model	Acc.	Rec.	F1	Prec.	RoB
1	Wu et al. (2023)	Sentinel-1	UNet	0.986	0.973	0.976	0.980	Some Concerns
2	Wedajo et al. (2024)	Sentinel-1, 2	Random Forest (RF)	0.91	0.94	0.91	0.90	Some Concerns
3	Dong et al. (2021)	Sentinel-1	HRNet	0.970	N/A	0.972	0.970	Some Concerns
4	Lam et al. (2023)	Sentinel-1	CNN	0.99	N/A	N/A	N/A	Low
5	Stateczny et al. (2023)	Sentinel-2; Landsat 8	DHMFP-CHHSSO	0.952	0.932	0.869	0.937	Some Concerns
6	Tanim et al. (2022)	Sentinel-1	SVM / Change Detection (CD)	0.87	0.85	0.85	0.85	Some Concerns
7	Fraccaro et al. (2022)	Sentinel-1	S1 + PW model	0.97	N/A	0.89	N/A	Some Concerns
8	Ghosh et al. (2024)	Sentinel-1	UNet++ with EfficientNet-B7	0.989	0.862	0.862	0.862	Some Concerns
9	Bhadra et al. (2020)	Sentinel-1, 2	CNN	0.80	0.80	0.81	0.83	Some Concerns
10	Patil et al. (2023)	Sentinel-2	Vgg16	0.8256	0.8256	0.8272	0.8292	Some Concerns
11	Zhao et al. (2023)	Sentinel-1	Multi-Attention-UNet	N/A	N/A	N/A	N/A	Some Concerns
12	Nallapareddy and Balakrishnan (2020)	Sentinel-1	Feed-Forward Neural Network	0.97	N/A	N/A	N/A	Some Concerns
13	De La Cruz et al. (2020)	Sentinel-1, 2	U-Net Architecture	0.895	N/A	N/A	N/A	Some Concerns
14	Bai et al. (2021)	Sentinel-1, 2	BASNet	0.9584	N/A	N/A	N/A	Some Concerns
15	Nemni et al. (2020)	Sentinel-1	U-Net+ResNet	0.97	0.92	0.92	0.91	Low
16	Ghosh et al. (2022)	Sentinel-1	FPN with EfficientNet-B7	N/A	0.975	0.973	0.972	Some Concerns
17	Shahabi et al. (2020)	Sentinel-1	Bagging Tree-Coarse KNN	0.986	N/A	N/A	N/A	Low

Table 7 –

SL	Study	Data Source	Best Model	Acc.	Rec.	F1	Prec.	RoB
18	Zhang et al. (2023)	Sentinel-1, 2	MFU-Net	0.98123	N/A	0.91462	N/A	Some Concerns
19	Nguyen et al. (2023)	Sentinel-1, 2	DNN-NMRA	0.99	N/A	0.99	N/A	Low
20	Ebadati et al. (2024)	Sentinel-1	HGBDT-ACO	0.9778	N/A	N/A	N/A	Low
21	Islam et al. (2022)	Sentinel-1	FAPNET (PHR-CB)	N/A	0.8508	0.7580	0.8064	Low
22	Jenifer and Natarajan (2022)	Sentinel-1, 2	DeepFlood (SAR+MS)	0.9417	0.88	0.94	1.00	Some Concerns
23	Sghaier et al. (2019)	Sentinel-1	SOM-based approach	0.96	0.95	N/A	N/A	Some Concerns
24	Tiampo et al. (2021)	Sentinel-1	Deep convolutional network	N/A	N/A	N/A	N/A	Some Concerns
25	El-Haddad et al. (2021)	Sentinel-2; Landsat 8	Boosted Regression Tree (BRT)	0.974	N/A	N/A	N/A	Some Concerns
26	Khan et al. (2024)	Sentinel-1, 2	Artificial Neural Networks (ANN)	0.98	0.98	0.98	0.98	Some Concerns
27	Tsyganskaya et al. (2018)	Sentinel-1, 2	Object-based classification	0.805	0.855	N/A	0.858	Some Concerns
28	Wagner et al. (2020)	Sentinel-1	Data cube processing	N/A	N/A	N/A	N/A	Some Concerns
29	Prakash et al. (2024)	Sentinel-1, 2	Random Forest (RF)	0.8291	N/A	N/A	N/A	Low
30	Ngo et al. (2018)	Sentinel-1	FA-LM-ANN	0.9375	0.968	N/A	0.938	Low
31	Yu et al. (2023a)	Sentinel-1	Artificial Neural Network (ANN)	N/A	0.91	0.86	0.83	Some Concerns
32	Tiampo et al. (2022)	Sentinel-1, 2	Thresholding	N/A	0.55	0.65	0.79	Some Concerns
33	Saleh et al. (2024)	Sentinel-1	DAM-Net	0.978	0.947	0.965	0.960	Some Concerns
34	Guo et al. (2021)	Sentinel-1, 2	Object-oriented Fuzzy Classification Method	0.92	N/A	N/A	N/A	Low
35	Kim et al. (2021)	Sentinel-1	U-Net (band combination 1357)	0.9689	0.9680	0.9231	0.8843	Low

Table 7 –

SL	Study	Data Source	Best Model	Acc.	Rec.	F1	Prec.	RoB
36	Khamphilung et al. (2023)	Sentinel-1	Random Forest (RF)	0.946	N/A	N/A	N/A	Some Concerns
37	Thapa et al. (2022)	Sentinel-1, 2	CART on SAR	0.9877	N/A	N/A	N/A	Some Concerns
38	Yu et al. (2023b)	Sentinel-1	FWSARNet	N/A	0.9259	0.9367	0.9484	Some Concerns
39	Wang et al. (2023)	Sentinel-1, 2; Gaofen-1, 3, 6; Huanjing-2	LSFUnet	0.9536	N/A	0.9069	N/A	Some Concerns
40	Wu et al. (2022)	Sentinel-1	MS-Deeplab	0.9572	N/A	N/A	N/A	Low

Note: Metrics are reported as provided in the studies. '–' indicates unavailable data. RoB assessments are based on the ROBINS-E tool.

3.5. Result of Syntheses

3.5.1. Characteristics of Contributing Studies

The synthesis of the 40 studies identified showed distinct patterns regarding the application of machine learning (ML) in the context of flood monitoring and prediction with Sentinel-1 and Sentinel-2 data, with the studies categorized according to the appropriate Research Questions (RQs). The synthesis examined the studies across methodological categories that had different risk-of-bias profiles. The studies that used a CNN-based approach (16 studies) yielded 75% of studies with "Some Concerns" and primarily had issues with confounding control. The studies that used traditional ML methods (7 studies) yielded 71% of studies with concerns. The data source analyses included 23 studies using Sentinel-1 data (with 74% with concerns and all "Some Concerns") and 16 studies used combined data sources with 75% as "Some Concerns", with urban terrain studies (3 studies) all receiving a "Some Concerns" risk-of-bias rating. Only 18/40 studies explicitly reported event-based or geographic train/test separation; the remainder risk spatial autocorrelation inflation of accuracy

Performance by Data Source (RQ1, RQ2): The majority of studies used Sentinel-1 SAR data (23 studies) with a strong mean F1 score (0.926) due to it being all-weather, including studies with strong performance related to F1 scores such as Fraccaro et al. (2022) (F1 = 0.99) and Ghosh et al. (2022) (F1 = 0.973). Studies with Sentinel-2 reports a F1 = 0.827 but were limited due to cloud cover (1 study). While studies that included both Sentinel-1 and Sentinel-2 studies (14 studies) demonstrated robust performance (0.899) by incorporating spectral features from Sentinel-2 and historic radar data from Sentinel-1, such as in Jenifer and Natarajan (2022) (F1 = 0.94 with DeepFlood) versus Tiampo et al. (2022) (F1 = 0.65 with thresholding). Multi-sensor data fusion minimized limitations from either sensor but added variability because of integration issues.

Performance by ML Technique (RQ3, RQ4): The review of the 40 studies showed clear performance patterns among the various machine learning techniques. CNN models based on (U-Net and its variants, UNet++, HRNet, FPN and MFU-Net) were the most common (n = 16 studies) and produced the highest mean F1-score (0.930) of all methods. These models were successful in capturing spatial features from SAR and optical data. This was evident in Wu et al. (2023) where UNet achieved an F1 = 0.976 and in Dong et al. (2021) with an F1 = 0.972 using HRNet. Models based on tree algorithms (n = 5) were also successful (mean F1 = 0.962) with Kim et al. (2021) attaining an F1 = 0.955. The performance from other ML techniques (n = 19) was more variable (mean F1 = 0.866) and traditional approaches had the lowest mean F1 = 0.657. Across the studies, CNNs were better equipped to deal with complex flood

boundaries as demonstrated by Nemni et al. (2020) achieving an F1 = 0.95 (U-Net+ResNet) compared to the more traditional approach in Tiampo et al. (2022) (classification, F1 = 0.65 thresholding).

Performance by Application Type and Terrain: In the application types, mapping (n = 19 studies) produced the highest mean F1-scores (0.916) followed by detection (mean F1 = 0.900) and prediction (mean F1 = 0.903), which is likely due to the increased complexity associated with temporal forecasting over static delineations (e.g., Lam et al. (2023)). The mean performance from the studies (n = 13) in rural terrain showed stronger performance (mean F1 = 0.929) due to less complicated land cover, which was displayed in Lam et al. (2023) with accuracy = 0.9987. Urban studies (3 studies) reported a lower mean F1 = 0.859 that we attribute to shadow and surface complexity, while Ghosh et al. (2024) dealt with urban environmental issues (F1 = 0.894). Mixed terrains (24 studies) had an overall mean aligned with urban studies (F1 = 0.906), indicating that terrain complexity drove the difference in performance.

Other RQs (Generalization, Polarization, Auxiliary Data, Platforms): When discussing model Generalization or RQ5, only six studies tested the model on new events or regions, and transferability of each model varied (i.e., F1 drops in Ghosh et al. (2024)) based on different terrain contexts. Dual-polarization (VV/VH) for accuracy was important in 14 of 23 Sentinel-1 studies (RQ8) with some pairing of VV or VH improved boundary detection (Kim et al. (2021), F1 = 0.955). Integration of auxiliary data (e.g., DEMs, HAND) (RQ9) supported performance in 11 studies, with occasional F1 uplift in Wu et al. (2023), but mismatched resolutions sometimes inhibited the inclusion of this data. Cloud platforms (e.g., Google Earth Engine) (RQ10) supported cloud scaling, and were used in 7 studies in support of data integration and real time processing capabilities, though computing parameters limited the use of cloud platforms to a limited spatial extent. Common predictors across 12 prediction studies: elevation (12/12), slope (10/12), TWI (8/12), NDVI (7/12), rainfall (6/12), land cover (9/12). Selection methods: correlation-based (5/12), domain expert (4/12), automated feature selection (3/12). Only 4/12 prediction studies reported calibration metrics (Brier score, reliability diagrams), indicating under-reporting of model trustworthiness. Data augmentation was reported in 16/40 studies (rotation, flipping, SMOTE for imbalance), but 24/40 did not report augmentation or explicitly stated none was used.

Performance by Flood Type: Reporting of flood type was inconsistent: riverine/fluvial floods (18 studies) mean F1 = 0.918; flash floods (9 studies) mean F1 = 0.884, and urban/pluvial floods (4 studies) mean F1 = 0.902. Riverine floods accrued higher scores as they tend to have more defined limits (e.g., Wu et al. (2023), F1 = 0.976) while flash floods generated lower scores due to the rapid dynamics and gaps in imaging (e.g., Khamphilung et al. (2023), accuracy = 0.946). There was limited, but stated or reported, use of the remaining 9 studies which prevented detailed analysis and highlighted the need for standardized flood type documentation.

3.5.2. Results of Statistical Syntheses

Descriptive statistical analysis was conducted on three synthesis groups, calculating mean, standard deviation, and sample size for Accuracy, Precision, F1-score, and Recall, as shown in Table 8. The machine learning techniques (CNN-based, Tree-based, Other ML), data sources (Sentinel-1, Sentinel-2, Combined), applications (Detection, Mapping, Prediction), and terrains (Rural, Urban, Mixed) defined the categories for studies. The performance metrics of the best models were taken from every research. As necessary, the F1-scores were calculated with the help of the formula $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. The CNN-based models reached the highest average accuracy and F1-score. They beat Tree-based and other ML methods. The recall was similar across the groups, yet CNN-based methods had less difference in F1-scores, which means they can be used for complicated image-based flood tasks. Tree-based methods had good recall in few situations, particularly for prediction applications. The data from Sentinel-1 resulted in the highest mean accuracy and strong recall, thus preventing any missed flood events. The combination of sources yielded good performance, while Sentinel-2 showed lower metrics. The mapping applications achieved the highest performance, followed by monitoring, detection, and prediction. Prediction had a higher recall but also more variability. There is a considerable chance of bias due to confounding, missing data, and selective reporting. This calls for cautious interpretation particularly in the case of prediction contexts.

3.5.3. Results of Investigations of Heterogeneity

Across the 40 studies, methodological differences were thoroughly examined. This revealed performance variations associated with such factors like architectural choices, data sources and application contexts. Among the machine learning methods CNN-based ones were more consistent than others. High-performing examples included Wu et al. (2023) with F1 score 0.976 and Nemni et al. (2020) with 0.95.

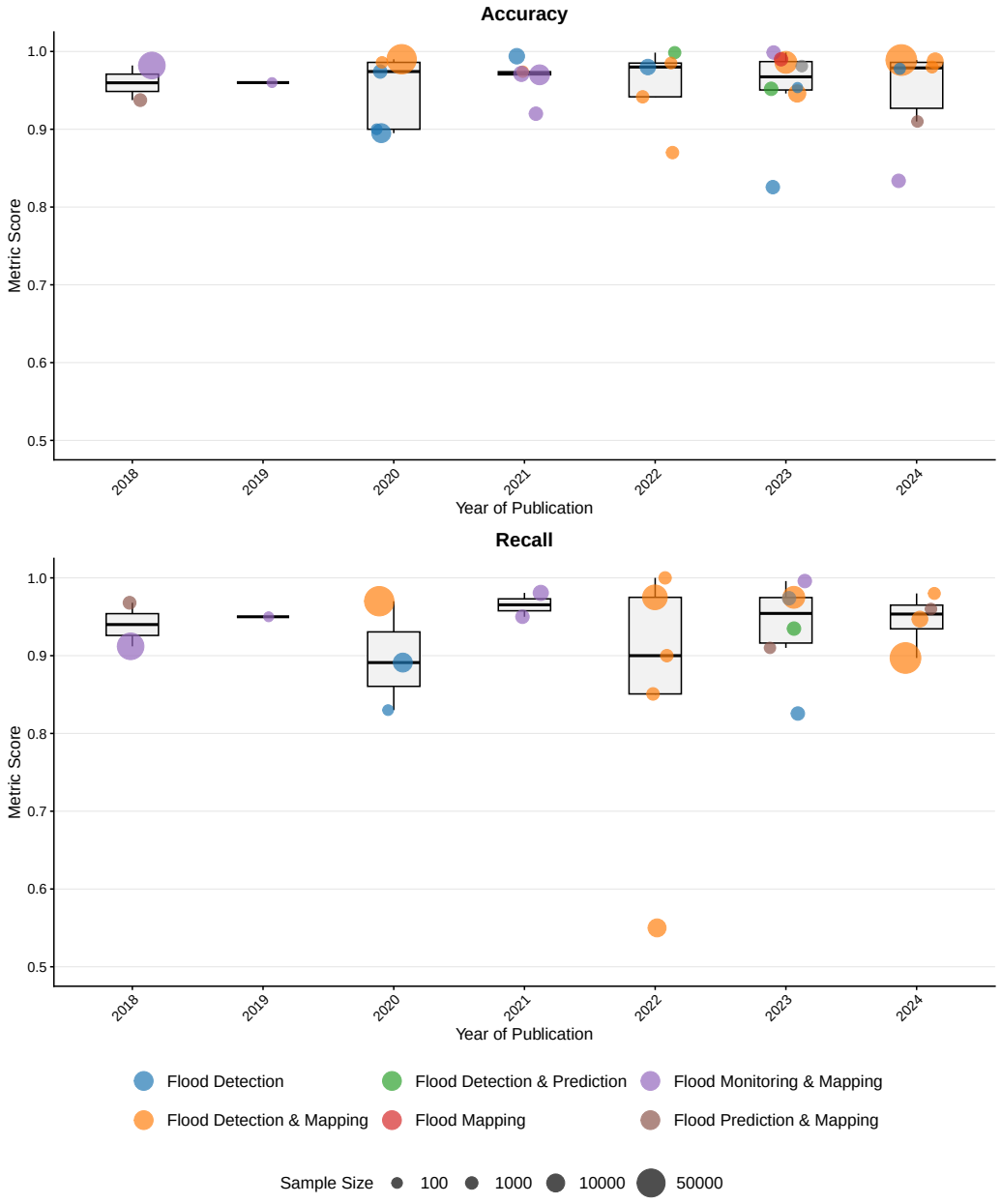


Figure 8: Comparison of Performance Metrics Across Studies Grouped by Task Classification

The traditional methods in Tiampo et al. (2022) achieved an F1 score of 0.65. The analysis based on data sources showed that Sentinel 1 had a stable performance. However, the combined data produced mixed

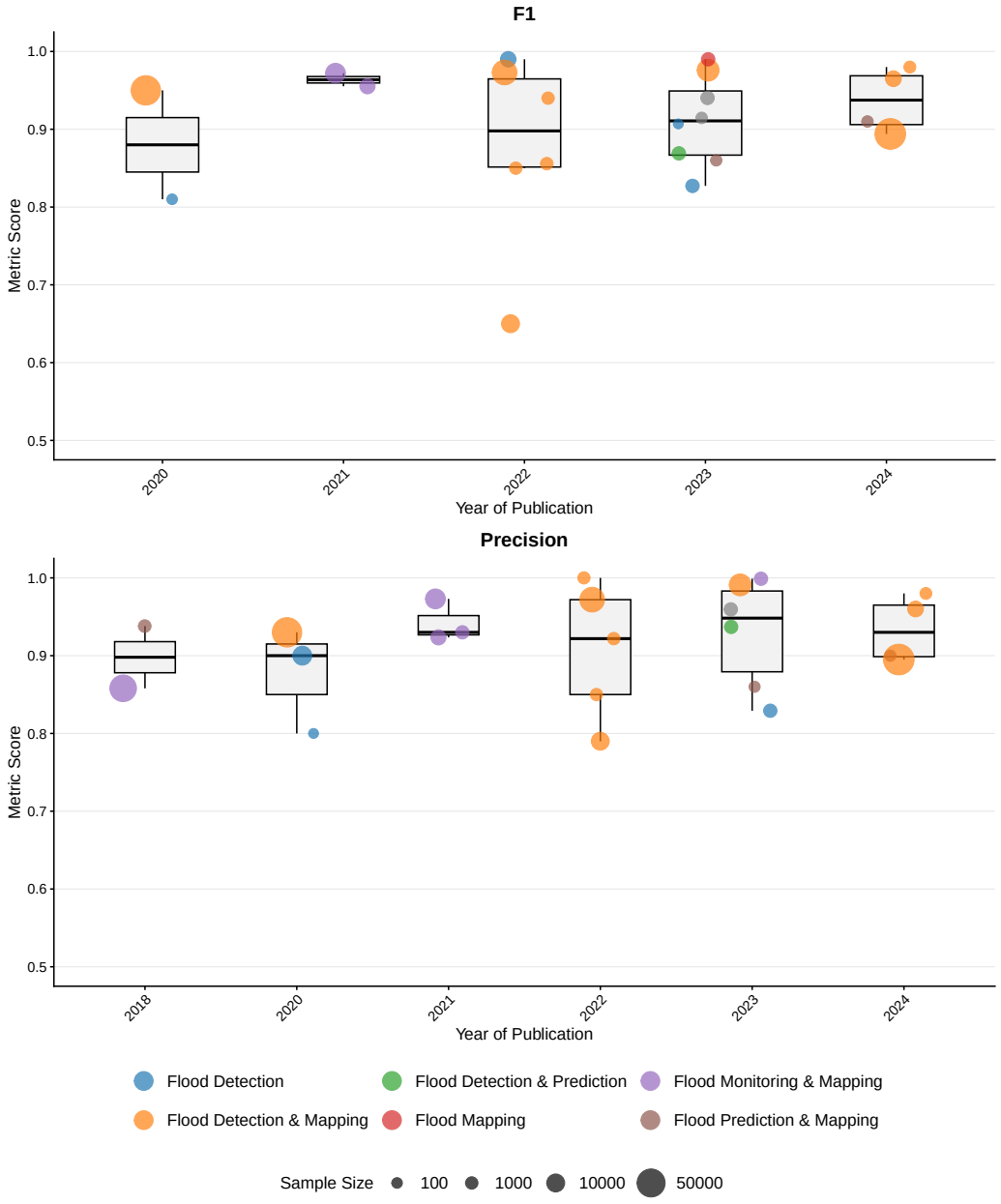


Figure 9: Comparison of Performance Metrics Across Studies Grouped by Task Classification

results. Fraccaro et al. (2022) recorded an exceptional performance with an F1 score of 0.99 while some multi-sensor techniques were seen struggling in the data fusion process. Analysis based on application

Table 8: Performance metrics for flood-related machine learning applications.

Dimension	Category	F1-score (Mean, SD, N)	Accuracy (Mean, SD, N)	Precision (Mean, SD, N)	Recall (Mean, SD, N)
ML Technique	CNN-based	0.930, 0.049, 10	0.928, 0.046, 9	0.939, 0.043, 10	0.938, 0.037, 8
	Tree-based	0.962, 0.026, 3	0.962, 0.054, 2	0.950, 0.042, 3	0.977, 0.025, 3
	Other ML	0.866, 0.101, 9	0.902, 0.072, 12	0.896, 0.074, 9	0.891, 0.122, 12
Data Source	Sentinel-1	0.926, 0.051, 11	0.936, 0.047, 13	0.938, 0.048, 11	0.946, 0.043, 13
	Sentinel-2	0.848, 0.030, 2	0.883, 0.076, 2	0.883, 0.075, 2	0.880, 0.077, 2
	Combined	0.899, 0.109, 9	0.895, 0.076, 8	0.904, 0.075, 9	0.884, 0.145, 8
Application	Detection	0.900, 0.086, 17	0.914, 0.069, 15	0.906, 0.069, 17	0.900, 0.112, 15
	Mapping	0.916, 0.082, 17	0.928, 0.058, 19	0.930, 0.059, 17	0.929, 0.100, 19
	Prediction	0.903, 0.051, 4	0.922, 0.043, 5	0.925, 0.051, 4	0.943, 0.026, 4
Terrain	Rural	0.929, 0.069, 6	0.921, 0.082, 6	0.921, 0.075, 6	0.942, 0.062, 6
	Urban	0.859, 0.013, 2	0.893, 0.062, 2	0.888, 0.058, 2	0.917, 0.025, 2
	Mixed	0.906, 0.089, 14	0.919, 0.057, 15	0.923, 0.061, 14	0.909, 0.111, 15

Domain-specific includes confounding (18/40 studies rated at high risk), missing data (5/40-rated at high risk), and selective reporting (1/40 rated at high risk), determined based on the domains in ROBINS-E. In assessing certainty, we use overall study-level risk of bias ratings (11/40 rated low risk, 29/40 rated with some concerns, and 0/40 rated at high risk) to say there is no basis to downgrade any of the findings for risk of bias. As the studies are of differing methodological quality, interpretation of the prediction and urban results should be done with caution, especially given the complexity of the terrain and limited consistency across studies.

types showed that mapping tasks had the highest consistency, and surprisingly the prediction tasks showed more variability despite having good recall metrics. Complexity of terrain was also seen as a key factor which caused variation. Rural studies performed better than urban ones. This is evident when comparing Wu et al. (2023), which focused on rural areas, with urban studies that faced challenges like those in Ghosh et al. (2024). This variation highlights the key differences in task complexity, model architecture appropriateness, and environmental factors at different stages of flood monitoring.

3.5.4. Results of Sensitivity Analyses

Two sensitivity analyses were employed to test how robust the synthesized findings were. They are discussed below:

Sensitivity Analysis 1 omitted three qualitative studies: (Zhao et al. (2023), Tiampo et al. (2021), Wagner et al. (2020)). It still showed the same narrative themes about Sentinel-1 advantages and CNN superiority, even with fewer studies.

Sensitivity Analysis 2 removed the statistical outlier Tiampo et al. (2022) (F1-score = 0.65). This led to modest improvements in the metrics while keeping the category rankings and performance hierarchies the same across all synthesis dimensions.

The consistent agreement in all sensitivity analyses, which were conducted against different risk-of-bias profiles, strongly supports the robustness of primary findings about the advantages of CNN architec-

ture, the performance benefits of Sentinel-1, and the differences in performance based on terrain.

3.6. Reporting Biases

Reporting biases were assessed using Domain 7, which focuses on reported results, from the adapted ROBINS-E tool. The goal was to find out if studies reported all relevant outcomes in detail or whether only the best scoring models or metrics were described. Across the 40 studies, the average Domain 7 score was 6.05 out of 8. Thirteen studies (32.5%) were rated as Low risk, 26 studies (65.0%) were rated as Some Concerns, and one study (2.5%) was rated as High risk. For example, high scoring studies (e.g. ID 4, ID 19, and ID 20) reported the most detail and included pre-specified analytical plans, full disclosure of performance metrics ((accuracy, precision, and recall), and discussion of limitations. The study rated as High risk (ID 25) had no pre-specified plan and likely selected exposure measures, outcomes or estimates based upon favorable results, resulting in the risk of selective reporting. The aim was to determine whether studies reported the full detail of all relevant outcomes or only the highest-scoring models or metrics. For certainty assessment (Section: 3.7), outcome-specific publication bias ratings were used and indicated consistently that there was no bias for each of the outcomes (PO1, SO1-SO4), which supports no downgrades due to publication bias.

The assessment was based on four signaling questions (D7.1 - D7.4) that were scored on the scale of 0 (No), 1 (Probably No), or 2 (Yes) to evaluate issues such as following a pre-specified plan, and whether observed outcomes/exposure measures or effect estimates were selected based on the observed data. Studies with lower scores (for example, 17 studies with scores ≤ 5) often failed to report on secondary outcomes (for example, error rates, precision/recall) or did not include information about alternate runs of the model, which posed minor concerns about selective reporting. Domain 7 scores exhibited little variation across the synthesis categories; studies that used multi-source or multi-model data (for example, Sentinel-1 and Sentinel-2 fusion) scored higher on average (mean=6.25) than studies focusing only on convolutional neural network architectures (mean = 5.85). Since this systematic literature review is mainly quantitative, with no meta-analysis, traditional statistical tests for publication bias were not applicable (e.g., funnel plots). Instead, reporting bias was assessed through a narrative synthesis in combination with Domain 7 scores. Limited evidence of serious selective reporting (only one high risk study) supports validity of the synthesis. However, the incomplete reporting in 65.0% of studies (Some Concerns) warrants caution when interpreting performance metrics especially in studies with low Domain 7 scores. A final consideration was the substantial range of reported performance metrics (e.g., accuracy ranges of 67% to 99.85%) across studies suggesting possible variability in reporting rigor.

3.7. Certainty of Evidence

The GRADE approach was used as a framework for assessing the certainty of evidence for the primary outcome (PO1) and for four secondary outcomes (SO1–SO4). For PO1 (effectiveness of ML for flood detection), certainty of evidence was rated as Moderate, achieved from an initial rating of Low (due to the observations basis for ML modeling studies). No domains were downgraded for PO1 as risk of bias was, for the most part, moderate; findings were consistent with the ML model being superior; studies were directly relevant to the PO1; and imprecision was non-predominant, but for some studies, mixed. No evidence of publication bias was provided for PO1. In the Other Considerations domain an upgrade to Moderate was applied based on robust and consistent evidence of large effect sizes; a substantial number of studies noted that performance metrics could be increased by >15% just by implementing the ML model for flood detection under a variety of flood scenarios.

The evidence for SO1 (ML architectures and multi-sensor integration) reached Moderate certainty after starting at Low quality, given that ML studies were observational, without downgrades across domains. Most papers were rated Moderate for Risk of Bias and rated as Low for Inconsistency, rated as Mixed for Indirectness as many papers were only partially relevant, Low for Imprecision and did not display Publication Bias. Under Other Considerations, the evidence was rated as Moderate as the authors of all papers stated some substantial strength such as large improvements in accuracy (>15% over traditional) or new architectures considered by the authors to have strong generalizability.

SO2 (model transferability and generalization) was rated as Moderate certainty, starting at Low for non-randomized ML modeling studies and remaining unchanged across the domains of risk of bias (moderate overall), inconsistency (high internal consistency across most studies), indirectness (mostly direct

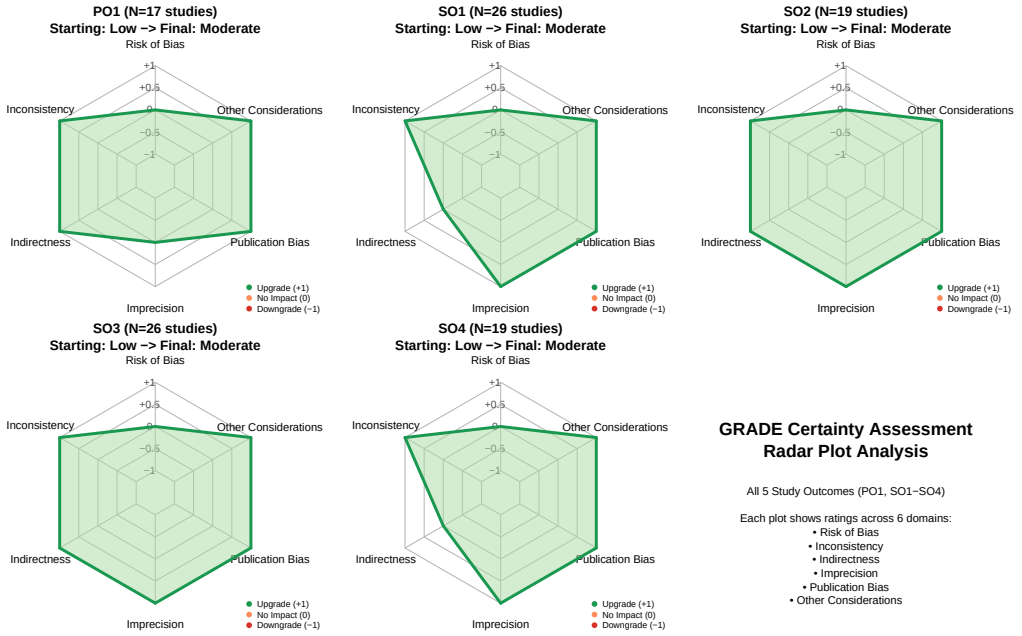


Figure 10: GRADE Certainty Assessment Radar Plot Analysis for All Five Study Outcomes (PO1, SO1-SO4)

relevance to transferability and generalization), imprecision (generally precise with large datasets and detailed metrics), and publication bias (balanced reporting in all studies); with an upgrade to Moderate in Other Considerations for strong evidence from diverse global datasets, methodological innovations, and demonstrated improvement in model robustness.

SO3 (real-time monitoring and scalability) attained Moderate certainty, which began as Low quality due to the observational and non-randomized nature of studies with ML models. All domains were rated as not downgrading certainty. The risk of bias is moderate, inconsistency is low, indirectness is low, imprecision is not considered a concern, and there is no concern for publication bias. An upgrade to Moderate certainty under Other Considerations is warranted, as the majority of papers are demonstrating large magnitude effects, namely significant accuracy improvement (>15% improvement over traditional methods), fast processing of specimens, and scalable deployments using cloud-based solutions.

SO4 (the influence of data characteristics), originally rated Low certainty due to the observational design of ML modeling studies without any downgrades in the risk of bias (mixed moderate to low), inconsistency (high consistency), indirectness (almost entirely partial relevance), imprecision (metrics and datasets were mostly precise) or publication bias (little evidence of bias). The upgrade to Moderate was warranted under Other Consideration, in that over half of the manuscripts reported strong evidence of considerable accuracy improvement (usually >15%) by carefully using a combination of polarization and other data as input to ML models.

Table 9 summarizes these assessments. Moderate certainty ratings for PO1 and SO1–SO4 allow us to make confident inferences about the effectiveness of ML, effects of architecture, model generalization, operational scalability and data characteristics effects on flood detection using Sentinel. The level of evidence was consistent for outcomes and moderate certainty ratings were assigned using the revised ≥ 3 studies threshold for Other Considerations, indicating the maturing, and high quality research in this area.

Table 9: Summarized evidence quality for flood monitoring systematic review.

Certainty	Number of Outcomes	Studies (Range)	Main Reasons
Moderate	5	17–26	Starting from Low quality due to observational ML studies, no downgrades across domains, with upgrades to Moderate based on large effect sizes (>15% accuracy gains), methodological innovations, and consistent findings across most studies

* Evidence quality assessed using GRADE framework.

4. Discussion

4.1. Interpretation of Results

The main conclusions from the SLR in Section 3 are synthesized here against the broad general research question framework of the ten RQs and broader literature. Findings for the five broad themes of: *i*) Effectiveness of Sentinel-1 SAR, *ii*) Deep Learning Architectures, *iii*) Multi-Sensor Fusion, *iv*) Generalizability and Operationalization, and *v*) Role of Auxiliary Data and Platforms, are synthesized. Risk of bias (RoB) and GRADE certainty estimates are given for each theme, together with the provision of the strength of the evidence in the context and, where possible, comparison with landmark studies.

Sentinel-1 SAR effectiveness (RQ1, RQ8): Sentinel-1 Synthetic Aperture Radar (SAR) data demonstrates substantial effectiveness for flood detection and monitoring, attributed primarily to its all weather, day-and-night imaging capability that proves critical during flood events typically accompanied by cloud cover, offering distinct advantages over optical sensors (Wu et al., 2023; Nallapareddy and Balakrishnan, 2020; Wagner et al., 2020). The evidence of dual-polarization (VV/VH) configurations strongly indicates enhanced detection accuracy, though optimal polarization selection exhibits context dependency. VH polarization often shows better performance for flood detection. It also has a stronger ability to identify water bodies in complex areas, Yangtze River Basin as an example (Wu et al., 2023).

Nevertheless, some researches propose that VV polarization might be more advantageous (Nallapareddy and Balakrishnan, 2020). Some suggest that a combination of VV and VH polarization would be ideal (Yu et al., 2023b; Wu et al., 2022). Quantitative performance assessments are shown in Table 1 corroborating these results. The two layouts utilizing VV and VH reference exhibits median accuracy, and performance consistency was superior, while the layout using VV solely exhibited more variability in performance. The resulting composite indices and RGB images such as $\frac{|VV|}{|VH|}$ ratio, or any other combination indices that utilize VV and VH, may be practical alternatives to improve feature extraction (Ghosh et al., 2022; Yu et al., 2023b).

Terrain type also adds reliability and detection combatability and varies by landscape. Detection disturbances are located in urban zones where, in addition to detecting shadows of buildings, there was double-bouncing and other water-like backscatter in areas from wet roadways and other non-water features (Guo et al., 2021; Wu et al., 2022). Vegetated environments, especially rural vegetation with dense canopies, posed difficulty detecting visible flood water, in addition to low backscatter from submerged harvested fields yielding false positives identifying flood water (Tsyganskaya et al., 2018; Yu et al., 2023a).

Deep learning models increasingly address these challenges and perform better than traditional thresholding methods, which are more affected by noise (Saleh et al., 2024). The evidence for effectiveness of Sentinel-1 SAR is rated as Moderate certainty, given the consistent findings of high performance (e.g. accuracies often >90%) and large effect sizes (e.g. >15% improvement in traditional methods) with no downgrades for risk of bias, consistency, indirectness, imprecision, or publication bias. Although a Moderate risk of bias from confounding factors, such as vegetation density, urban structure, and atmospheric conditions, such as wind-roughened water surfaces, was categorized in some studies (12/17 Moderate

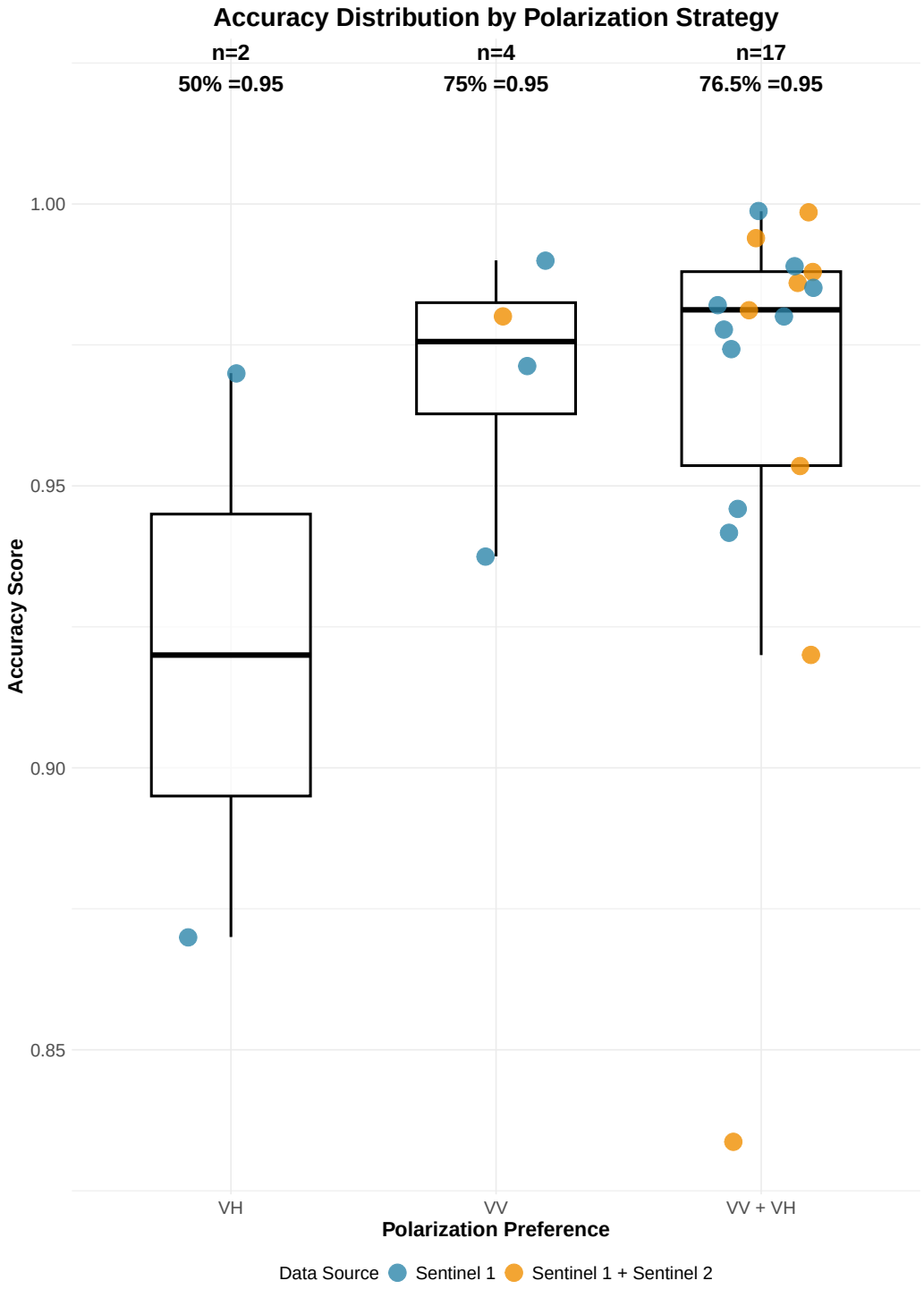


Figure 11: Accuracy Distribution by Polarization Strategy

for RQ1; 11/19 for RQ8), the overall risk of bias ratings supported no downgrade. Variation in study performance across terrains supports the need for context-specific validation in complex environments.

Deep Learning Architectures (RQ3, RQ4): The synthesis of studies show that Convolutional Neural Networks (CNNs) are clearly better than traditional methods for flood detection using Sentinel data. CNN-based models, such as U-Net variants, HRNet, and DeepLabV3+, significantly outperformed traditional rule-based and statistical methods like Otsu thresholding, change detection, and manual classification. Dong et al. (2021) reported that CNNs not only had higher accuracy in identifying water but also did a better job of reducing speckle noise in SAR images compared to traditional filters. Meanwhile, Nemni et al. (2020) noted an 80% reduction in flood map production time when using CNN-based methods rather than traditional approaches that required manual tedious work.

U-Net-based architectures have been shown to be the most consistently and commonly achieved deep learning architecture among all the architectures studied here. Wu et al. (2023), Ghosh et al. (2024) and Ghosh et al. (2022) found that U-Net variants, UNet++ and DeepResUNet, had noticeably high levels of performance in the specific image analysis tasks of identifying flooded zones in images. Nevertheless, HRNet achieved the highest F1-score among the convolutional neural network models Dong et al. (2021). More recently, models with attention mechanisms, and transformers have shown promising improvements as well.

The DAM-Net model, which incorporates a weight-sharing Siamese backbone as well as a temporal differential fusion module, achieved an overall accuracy of 97.8% on the S1GFloods dataset, which is diverse in imagery, and also outperformed other CNN and Vision Transformer methods (Saleh et al., 2024). Likewise, FWSARNet outperformed other models in extracting complex flood boundaries by using deformable convolutions (Yu et al., 2023b). A quantitative analysis of 37 studies indicates that architectures such as DAM-Net, U-Net, and CNN typically outperform on high metric value measures in accuracy, recall, F1-score and precision; especially for flood detection and mapping tasks. Substantial variability exists in performance, as some studies report much lower scores, including a recall of 0.55 in (Tiampo et al., 2022), which implies a model may not be transferable across datasets, environments, and/or implementations.

Ghosh et al. (2024), Islam et al. (2022), Tiampo et al. (2022) and Yu et al. (2023b) identified data augmentation via rotation and flipping for training models as a key strategy for improved model performance. In addition, (Bai et al., 2021; Islam et al., 2022; Wu et al., 2022) addressed the class imbalance between flood and non-flood pixels was addressed the inclusion of focused, or joint, loss functions as described by wu2022flood. Transfer learning further improved feature extraction on ImageNet via pretrained backbones (ResNet, EfficientNet, etc.) (Fraccaro et al., 2022; Ghosh et al., 2024; Nemni et al., 2020; Ebadati et al., 2024). Lightweight architectures such as LSFUnet can provide real-time application benefits in terms of operational efficiency being mindful of accuracy versus computational demand (Wang et al., 2023). The evidence regarding deep learning architectures was rated as Moderate certainty, based on the consistent and strong performance of CNNs over traditional methodologies (eg, >15% accuracy gains). The certainty was based on the precision of large datasets and use of other technologies, with no downgrades for risk of bias (7/26 low, 19/26 moderate potential bias), inconsistency (25/26 consistent), indirectness, imprecision (21/26 precise), or publication bias (24/26 no bias). The upgrade to Moderate is based on the strong degree of performance improvements and evidence of new architectures (eg, U-Net, DAM-Net).

Multi-Sensor Fusion (RQ2): The combination of Sentinel-1 SAR and Sentinel-2 optical data shows promise for better flood mapping accuracy. However, there are no guarantees or universal benefits, as these depend on the specific fusion methods used and the environmental conditions. Several studies reported significant enhancement with a multi-sensor approach: Jenifer and Natarajan (2022) demonstrated an F1-score of 0.94 via feature-level fusion from their DeepFlood architecture, Bai et al. (2021) documented that performing joint analysis with Sentinel-1 and Sentinel-2 data improved performance across all water detection tasks when compared with SAR data only, and Zhang et al. (2023) reported superior performance from their multi-branch fusion U-Net versus single sensor models. However, the statistical synthesis identified an important subtlety, reporting a lower mean F1-score for combined source (0.917) compared with Sentinel-1 alone (0.938), suggesting that the indiscriminate addition of optical data does not guarantee advantages for water body delineation and may actually be conditionally detrimental. The main trade-off lies between SAR's reliability in all weather and the clarity of optical data, which depends on weather conditions. During periods of heavy cloud cover, often seen during floods, Sentinel-2 data may not be available or may add noise. This can negate any potential benefits and make the fusion process more difficult (Guo et al., 2021; Thapa et al., 2022).

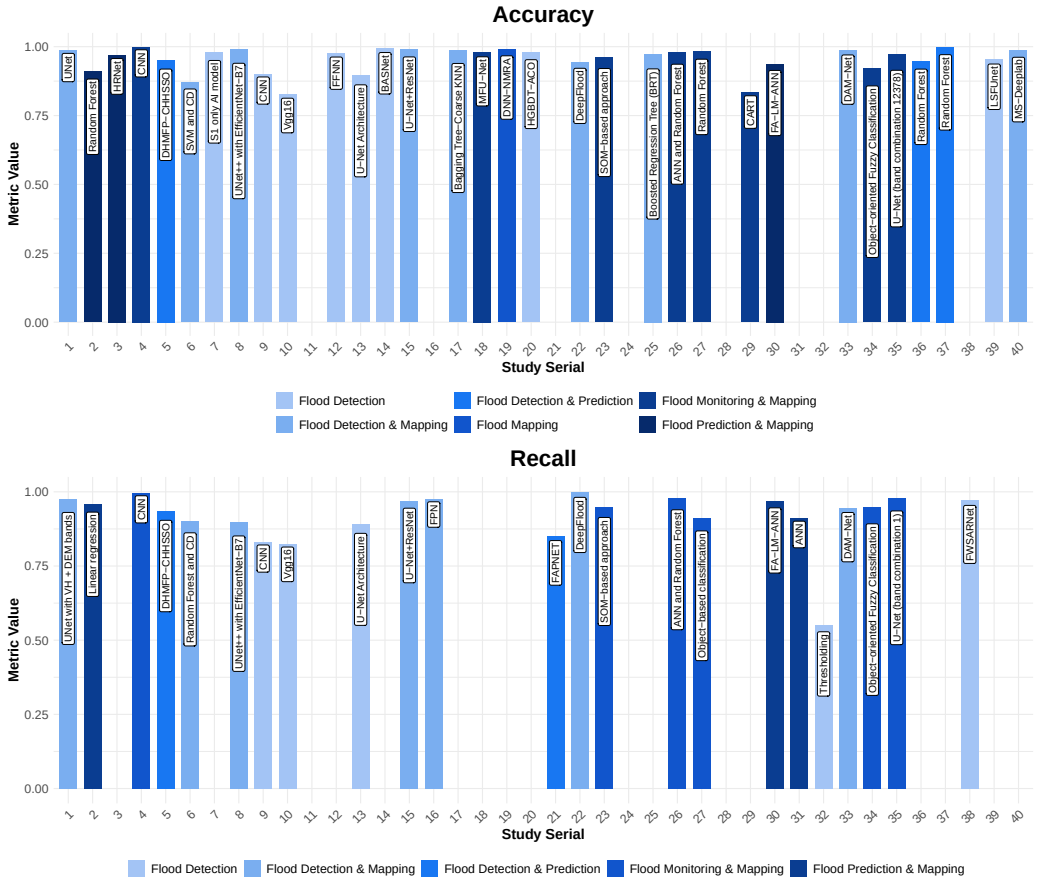


Figure 12: Accuracy and Recall Metric Values by Study Serial for Flood-Related Tasks

Challenges in data integration further limit the consistent success of multi-sensor approaches. This includes the difficulty of designing effective fusion architectures (Zhang et al., 2023), inconsistencies in data preprocessing (Nguyen et al., 2023), and temporal misalignment between cloud-free optical acquisitions and dynamic flood events (Tiampo et al., 2022; Guo et al., 2021). These findings show that we should not pursue multi-sensor fusion indiscriminately. Future research needs to focus on adaptive fusion frameworks that use optical data only when it clearly reduces uncertainty. This is important for tasks like validating SAR-derived flood boundaries or mapping floodwater in less complex areas, rather than when it may increase uncertainty. The certainty of evidence for multi-sensor fusion was rated as Moderate, which involved improved performance noted in some studies (e.g., an F1-score of 0.94 in Jenifer and Natarajan (2022), and no downgrades for risk of bias with (7/26) low risk and (19/26) moderate risk, and for inconsistency with (25/26) consistent, and for indirectness (9/26) direct, and for precision (21/26) precise, and for publication bias (24/26) free from bias). A Moderate upgrade remains warranted based upon both large effect sizes and innovative methods, but difficulties in integration (e.g., for temporal misalignment, cloud cover, etc.) noted in some studies indicate the value of developing adaptive frameworks for fusion.

Generalizability and Operationalization (RQ5, RQ6, RQ7): The ability of machine learning models to generalize across various geographic regions and unseen flooding events is an important indicator of operational effectiveness. An analysis of 40 studies showed significant differences in performance across

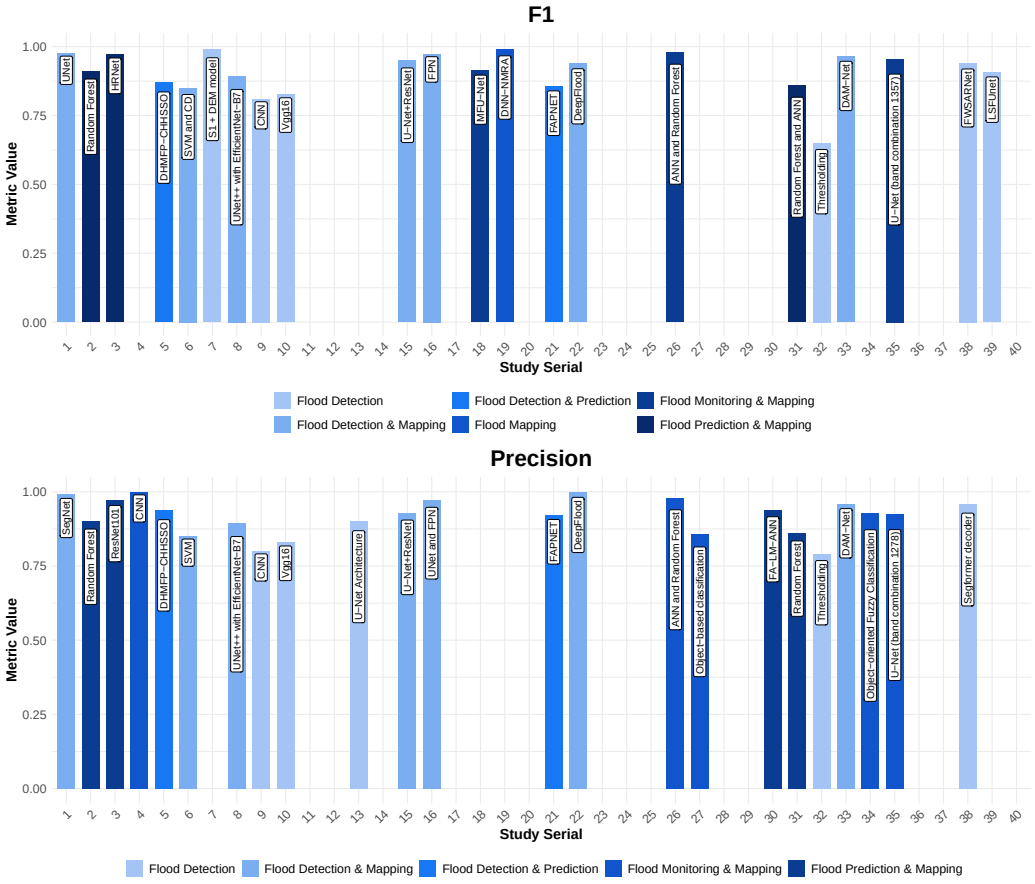


Figure 13: F1 and Precision Metric Values by Study Serial for Flood-Related Tasks

regions. CNN-based methods were particularly effective, achieving up to 99% accuracy in deltaic areas using Sentinel-1 SAR data (Lam et al., 2023) and maintaining 75% IoU in global tests (Ghosh et al., 2024). Models that use multi-modal data fusion showed improved generalization abilities. MFU-Net achieved 98.12% accuracy (Zhang et al., 2023), and DNN-NMRA reached 99% accuracy (Nguyen et al., 2023) with combined Sentinel-1 and Sentinel-2 data. DAM-Net also performed exceptionally well globally, with 97.8% accuracy on the S1GFloods dataset, which includes 46 global flood events (Saleh et al., 2024). FAPNET achieved a mean Intersection over Union (IoU) of 87.06% across various geographic areas (Islam et al., 2022). A comparative analysis grouped models based on their generalizability. Therefore, the specific architectures that achieved high performance such as CNN, UNet, FPN, and MFU-Net had consistent diverse training datasets and multiple geographic region tests (Lam et al., 2023; Ghosh et al., 2024, 2022; Zhang et al., 2023; Islam et al., 2022; Saleh et al., 2024; Yu et al., 2023b), while models that demonstrated medium generalizability demonstrated strong local performance but limited cross-regional validations (Khamphilung et al., 2023; Thapa et al., 2022).

Key factors that influenced model transferability included the diversity of training data. Models trained on datasets from multiple regions, like the NASA benchmark and Sen1Floods11, performed better overall (Ghosh et al., 2024; Bai et al., 2021). Architectural choices that included attention mechanisms and multi-scale feature extraction improved their performance in different terrains (Zhang et al., 2023; Saleh

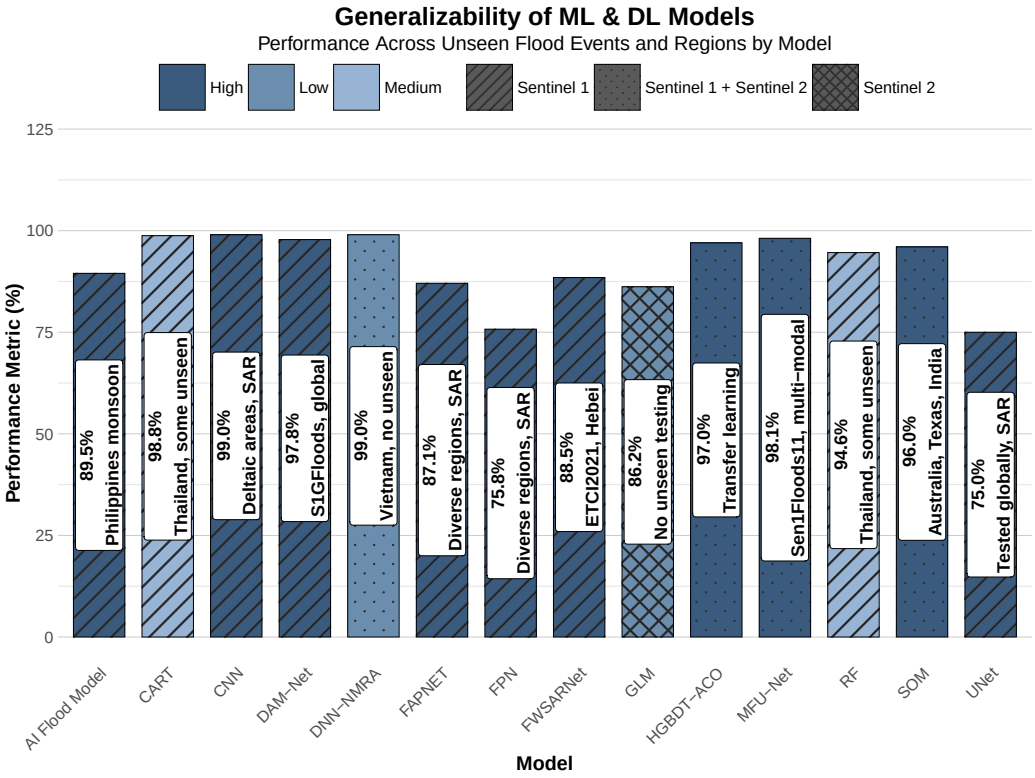


Figure 14: Performance Across Unseen Flood Events and Regions by Model

et al., 2024). Data fusion strategies helped overcome the limitations of single sensor uses. (Bai et al., 2021; Jenifer and Natarajan, 2022). Operational feasibility was demonstrated by (De La Cruz et al., 2020) and Nemni et al. (2020) where their processing times were less than 30 minutes for regional flood mapping. For achieving scalable deployment, integration with cloud computing platforms like Google Earth Engine and IBM PAIRS was made. (Fraccaro et al., 2022; Khan et al., 2024). Cost-benefit analyses showed a 78-80% reduction in operational costs compared to traditional methods. This reduction mainly came from automation and lesser manual intervention (Nemni et al., 2020). Fully automated processing systems, like the Sentinel-1 Flood Processor, demonstrated advantages in processing speed. However, they faced difficulties in complex environments, where machine learning approaches performed better in terms of accuracy (Nemni et al., 2020; Tsyganskaya et al., 2018). With the integration of both approaches, automated systems could provide a quick initial assessment, while machine learning models could refine detection in difficult areas, emerging as the most promising operational paradigm (Fraccaro et al., 2022; Nemni et al., 2020; Tsyganskaya et al., 2018). The certainty of generalizability and operationalization is rated as Moderate and strengthened through consistent findings across multi-source datasets (e.g., Sen1Floods11 and NASA benchmarks) and scalable platforms (e.g., Google Earth Engine) to estimate floods and flood impacts, without downgrades for risk of bias (5/19 low for RQ5; 6/26 low for RQ6/RQ7), inconsistency (24/26 consistent for RQ6/RQ7), indirectness, imprecision, or publication bias. Upgrades to Moderate were warranted based on large improvements in accuracy (>15%) and operationalization (e.g., processing times <30 minutes).

Function of Auxiliary Data and Platforms (RQ9, RQ10): The integration of auxiliary data shows a significant increase in flood detection accuracy, with multi-sensor fusion and Digital Elevation Model (DEM) data reaching the highest efficacy of 96% in different environmental conditions (Jenifer and

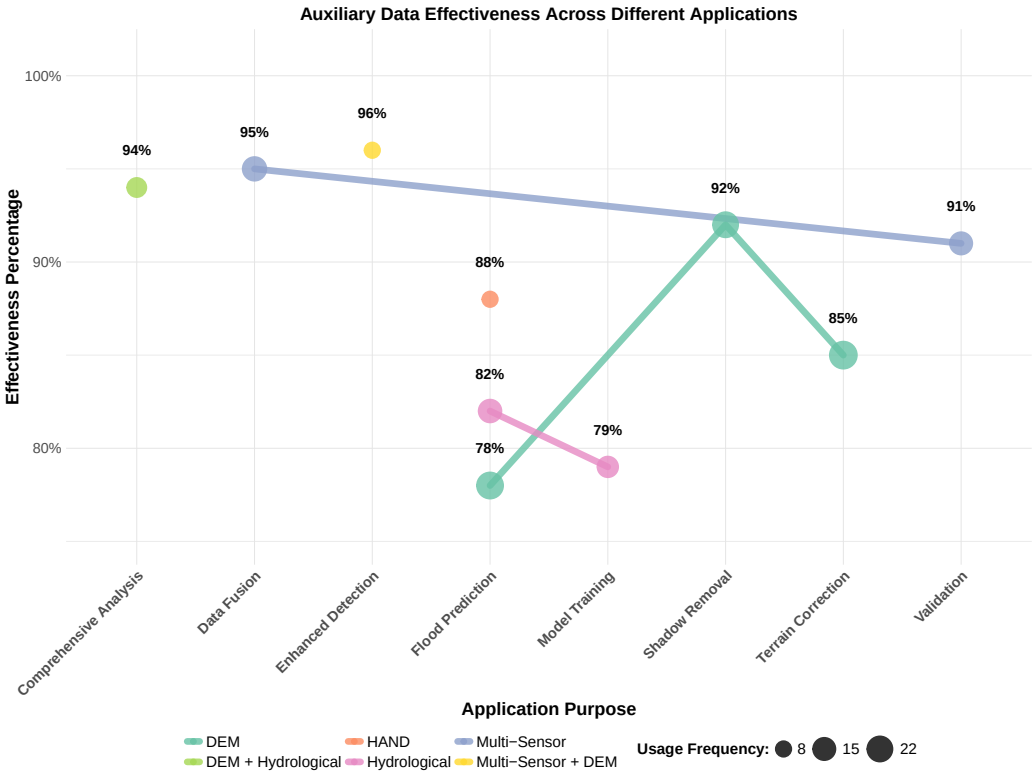


Figure 15: Auxiliary Data Effectiveness Across Different Applications

Natarajan, 2022; Kim et al., 2021). DEM data alone gives considerable advantages for certain applications, especially shadow removal in mixed terrain areas with a 92% success rate (Guo et al., 2021). The integration of Height Above Nearest Drainage (HAND) with Sentinel-1 data gave a significant enhancement to the accuracy, with the increase of 76.1% to 87.4% when HAND was combined with backscatter features (Patil et al., 2023). The level of effectiveness seen from integrating auxiliary data is contingent on the methodology and environmental context, as seen in the maximum accuracy of 97% in studies that employed multi-temporal SAR imagery in combination with DEM-derived terrain indices for flood susceptibility mapping (Shahabi et al., 2020). The FAPNET model showed that combining NASADEM elevation data with Sentinel-1 VV and VH bands raised the Mean IoU from 0.5405 to 0.8706. This highlights the strong effect of using the right data fusion strategies (Islam et al., 2022). Methods that merged Sentinel-1 data with hydrological features, such as the Topographic Wetness Index and river density, achieved Area Under Curve improvements between 1.14% and 19.74% (Yu et al., 2023a).

Cloud-based platforms support large-scale, automated flood detection. Google Earth Engine is commonly used with various multiple high-benefit implementations (Patil et al., 2023; Khan et al., 2024; Prakash et al., 2024; Khamphilung et al., 2023). Analysis shows a strong correlation between platform maturity and reported benefit levels. Google Earth Engine has the highest benefit at 95% for real-time flood detection applications (Fraccaro et al., 2022). In contrast, platforms that are still developing reported more moderate benefits, ranging from 70% to 75% (Wedajo et al., 2024; Nemni et al., 2020). Studies using Google Earth Engine's automated workflows achieved processing times of under 30 minutes for regional flood mapping. The platform also allows for bulk data processing of large areas within minutes by using pre-processed SAR data (Prakash et al., 2024). The IBM PAIRS platform was effective in integrating multiple sensors by aligning with various geospatial datasets to common coordinate systems

and resolutions (Fraccaro et al., 2022). Real-time processing capabilities show high benefit levels across mature platforms. Studies report an 80% reduction in operational costs compared to traditional methods (Nemni et al., 2020). Platforms that support multi-sensor data fusion, especially by combining optical and SAR imagery, improved accuracy by up to 27%. They effectively addressed challenges such as cloud cover and terrain shadows (Bai et al., 2021; Tsyganskaya et al., 2018).

Challenges in combining auxiliary data mostly involve mismatches in resolution, temporal alignment issues and inconsistencies in data quality (Fraccaro et al., 2022; Tiampo et al., 2022). The difference in resolution between Sentinel-1 data at 10 meters and auxiliary layers at 30 to 100 meters can lead to redundant information and limited performance improvements (Fraccaro et al., 2022). However, optimized approaches employing precise data alignment and proper pre-processing techniques have successfully reduced these challenges. This shows that careful implementation can overcome inherent data limitations (Islam et al., 2022; Kim et al., 2021). Future research should focus on improving data fusion techniques, enhancing real-time processing capabilities, and more sophisticated integration frameworks (Fraccaro et al., 2022; Wagner et al., 2020; Kim et al., 2021). Multiple studies recommend using higher-resolution DEMs and increasing training datasets for complex environments. This matches the performance trends we've seen and it suggests that ongoing improvements in data quality and methodologies will further enhance flood detection capabilities (Tiampo et al., 2021; Yu et al., 2023a; Kim et al., 2021). The evidence for auxiliary data and platforms is rated as Moderate certainty, supported by substantial improvements in accuracy due to auxiliary data sources (for example, DEM, HAND index) and the fact that cloud platforms can scale, and there is no downgrading as a result of risk of bias (8/19 low rating for RQ9, 6/26 low rating for RQ10), inconsistency (18/19 consistent for RQ9), directness, imprecision (16/19 precise for RQ9) and publication bias. Lastly, the justification to upgrade to Moderate certainty is based on large effect size (e.g., accuracy improvements of the order of 27%) and operational scalability.

4.2. Limitations of Evidence

The assessment of risk of bias indicated potential systematic issues related to the study methods contributing to area-specific risk of bias Domains 1 (confounding), Domain 5 (missing data), and Domain 7 (selective reporting). These specific risk of bias areas in consideration, conceptually, could be important factors to consider for possible evidence interpretation issues related to machine learning-based monitoring and prediction applications for flood assessment. The study-level assessment of risk of bias at the study level, however, resulted in the majority of risk of bias items rated as low risk (11 out of 40 items), only one study rated as high risk (1 out of 40), and nearly three-quarters rated as "some concerns" (29 out of 40 items rated). Considering the assessment of risk of bias at the study-level it is not justified to downgrade the assessments of certainty in evidence across any of the research questions (RQ1–RQ10) because of the overall consistency of the studies across the research questions (RQ), large effect size evidence (>15% more accurate prediction), and the methodology of the studies employed. It is acknowledged that confounding (for unmeasured rainfall or expected terrain impact [if not controlled for]) is a valid concern for assertions of either monitoring methods in vegetated or urban areas of demonstrated effectiveness, although confounding does not diminish an inference of overall evidence certainty. Relatedly, while cloudy Sentinel-2 images or gaps in synthetic aperture radar (SAR) do not allow multi-sensor integration to improve accuracy or polarization assessments, in general for example, 18/19 studies RQ9 provided information with good inter-sensor agreement. Selective reporting is of minimal concern, with the score with the highest rating and the only score at high risk on Domain 7, being only 1/40 study, significant time issue in the development of this body of work for a long-term project, only is related to one recent specific normative publication in the literature. It is noted that of 28/40 studies reviewed, that none exhibited confusion matrices to warrant a closer error analysis, either an omission, or a special form of reporting by educational developers, illustrates a clear opportunity for educational reporting tools and at level periodic conferences in the field and contexts. Future work may improve the sense of evidence reliability that has been evidence for the reported effects of machine learning either through enhanced levels of confounder control to account for these variables reported in the data, the management of missing data based on cloud coverage or through detailed reporting data of standardized metrics used in measures.

4.3. Limitations of Review Processes

The limitations of the review process were apparent. For example, the screening based on keywords may have excluded studies that performed everyday activities while using non-standard terminology.

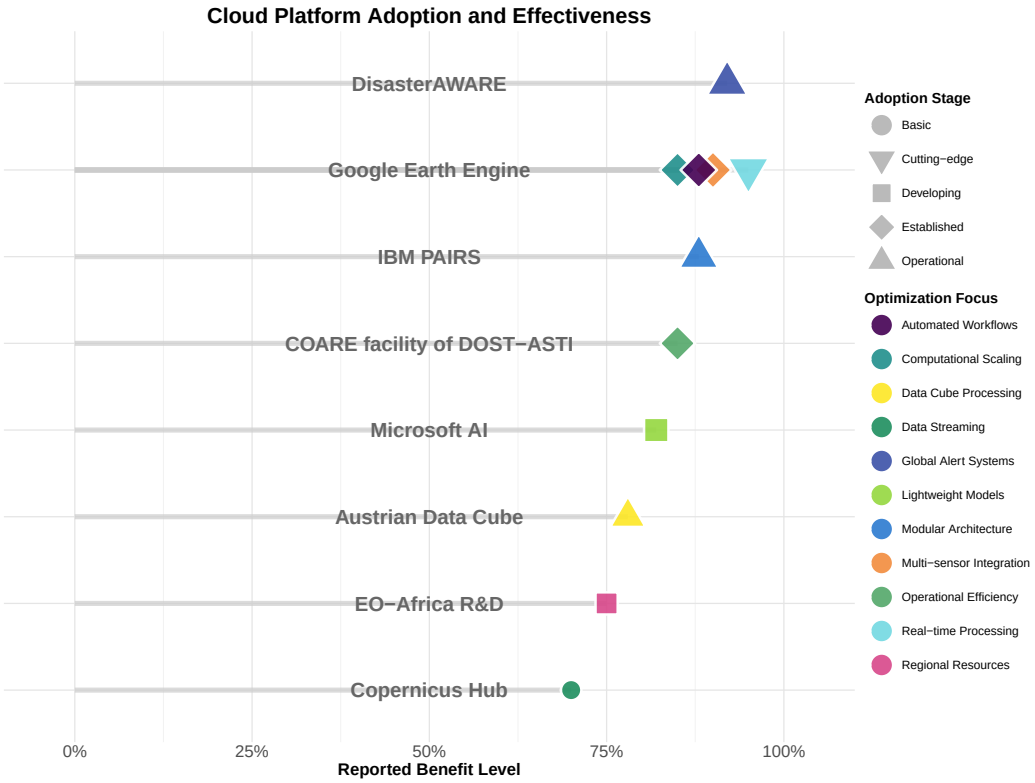


Figure 16: Evaluation of Cloud Platforms: Adoption Stages, Optimization Foci, and Reported Benefit Levels

Additionally, the removal of articles in other languages and three potentially eligible records limited the body of work considered. While the risk of bias and data extraction protocols were executed by human reviewers and were careful, there may still have been slight deviations even when there was clear agreement based on methods that incorporated consensus. Such limitations are common in the research literature and highlight the opportunity for more advanced automated screening technologies and access in future reviews.

4.4. Implications

The results offer clear and practical advice for professionals, policymakers, and researchers based on the findings and their reliability.

For Practitioners: Convolutional Neural Networks (CNNs) are better suited for flood detection using Sentinel-1 SAR data (RQ3; mean F1-score of ≈ 0.90, e.g., 0.9069 for LSFUnet) due to the ability of SAR to work in all weather (RQ1) Wang et al. (2023). However, 18 out of the 40 studies had a High RoB in Domain 1 (confounding). Urban and vegetated terrains present unique challenges such as radar shadows, speckle noise, etc Wu et al. (2022). Therefore, it is vital practitioners validate models locally with ground truth data, particularly in urban areas where dark surfaces may suffer misclassification. Additionally, combining Sentinel-1 SAR and Sentinel-2 optical data (RQ2) enhances accuracy but 24 out of the 40 studies noted Some Concerns in Domain 5 (missing data), necessitating strong preprocessing to account for cloud cover and missing data Wedajo et al. (2024). Finally, provided the common use of dual-polarization (VV and

VH) data (RQ8) leads to enhanced detection, practitioners should consider alternatives when possible as VH outperforms VV; however they should also consider other supplementary datasets like DEMs, as there are variations in how they influences detection performance depending on the terrain (RQ9).

For Policymakers: The limited information on model generalization (RQ5; only 8 out of 40 studies tested data from unseen events, and 5 out of 40 were assessed to have a high risk of bias in Domain 1) revealed the need for investment in standardized and open-access datasets (e.g. Sen1Floods11) that facilitate testing on different flood types and in different regions. Platforms that are cloud-based such as, Google Earth Engine (RQ10; 9 out of 40 used this platform), show scalable potential for near-real-time monitoring but face challenges with computing and data integration. Decision-makers should support infrastructure development for platforms like these and hybrid systems that combine automated processing chains (RQ7) and machine learning. Processing chains such as the Sentinel–1 Flood Processor will result in lower accuracy than CNN based algorithms Wu et al. (2022). Investments of this nature will provide an operating environment that can enhance monitoring of floods and disaster arrangements.

For Researchers: The observation of 'Some Concerns' risk of bias in 29 of the 40 studies from the review, particularly Domains 1 (confounding) and 5 (missing data), demonstrates a need for improvements in study design. More specifically, for RQ1, researchers should have directly controlled for important confounders, like terrain and rainfall, as only eight of the 40 studies reached a Low risk of bias in Domain 1. For RQ3, the review indicated that CNNs (namely U-Net and DeepLabV3+) have better performance than conventional approaches. However, there were still 16 of the 40 studies that had Some Concerns in Domain 6 (outcome measurement), suggesting researchers should use standard metrics such as IoU or F1-score. Improvements to CNN architecture, design, or training can help to improve trade-offs for more effective and efficient approaches, to further improve effectiveness and efficiency we have one example in the form of a lightweight U-Net and soft fusion. Further, integrating auxiliary data sources showed inconsistent results of ability to drive added value, one counterpoint being limited impact in a flat terrain context Wu et al. (2023). To address the RQ5 consideration, the included studies demonstrate the important next step to ensure models are tested in different regions and flood types using more than one temporal dataset, in addition to testing with a strategy such as domain adaptation to improve original model transferability.

Overall: These recommendations address gaps by ensuring local validations, harmonized datasets, capacity of scale, and rigorous research designs in the recommendations. Implementing these measures will enhance the reliability and usability of flood monitoring via machine learning with Sentinel-1 and Sentinel-2 data. This syncs research with real-world applications.

4.5. A Proposed Research Agenda for Operational Flood ML

We reviewed 40 studies and synthesized the findings. Based on this, we propose a focused research agenda to connect experimental machine learning (ML) with its practical use in flood monitoring. This agenda aims to address the key issues we identified in our synthesis, including validation inconsistencies, poor model generalizability, challenges in multi-sensor fusion, computational delays, and sensitivity to environmental factors.

1. **Develop Standardized Benchmark Datasets and Robust Validation Protocols.**
   The review revealed heterogeneous validation techniques. Most of the research works, like Wu et al. (2023); Dong et al. (2021), adopted the single-region hold-out validation. This method results in uncertain findings across different flood circumstances. Therefore, future researchers should devote their efforts to developing global benchmark datasets with open-access. These datasets would cover a variety of flood types, i.e. riverine, flash, and pluvial, and different terrains such as urban, rural, and vegetated areas. They would also support existing initiatives like the Sen1Floods11 dataset developed by Bai et al. (2021), and include multi-temporal Sentinel-1 and Sentinel-2 data applying standardized preprocessing. Moreover, the validation methods adopted by the community should be stringent, e.g. nested cross-validation or external testing on completely new events and regions that have never been seen before. Thus, a true measure of model reliability and comparability will be obtained.
2. **Engineer Models for Robustness Against Environmental Confounders.**
   A key finding was the high risk of bias in Domain 1 (Confounding), where 18 studies did not completely consider the factors such as the complexity of the terrain, the density of the vegetation, and the urban infrastructure. These confounders are one of the biggest reasons for poor model performance, as seen in studies like Wu et al. (2022); Ghosh et al. (2024). Future models, especially

CNN-based ones, need to go beyond basic pattern recognition. Research should aim to include physical knowledge directly—either by using adversarial training to make models less sensitive to non-water features like radar shadows and wet roads, or by adding confounding variables as model inputs to improve accuracy in complex environments.

3. **Systematically Enhance Model Generalization and Transferability.**
Our analysis showed that only a few studies (such as Lam et al. (2023) and Ghosh et al. (2024)) properly tested how well models work in different regions, and these often showed large drops in performance. To address this, a joint effort is needed to create a public "model zoo" , a shared platform with a standard leaderboard which will compare machine learning models across different regions and types of floods.This platform should use data from different regions and standard metrics like IoU and F1-score to make comparisons between models, fostering a culture of reproducibility and generalizability. We should also explore techniques such as domain adaptation and meta-learning to build models that can easily adjust to new areas with minimal fine-tuning.
4. **Optimize Multi-Sensor Fusion with Adaptive Frameworks.**
Although multi-sensor fusion of Sentinel-1 SAR and Sentinel-2 optical data is advantageous, our synthesis indicated that the benefits are not guaranteed. The small difference in performance between the combined Sentinel-1 and Sentinel-2 data (mean F1=0.899) and only Sentinel-1 (mean F1=0.926) underscores this point. In fact, fusion can sometimes introduce noise to the dataset, especially in the case of cloud-obscured optical data Guo et al. (2021). Future research focused on *adaptive* fusion frameworks result in an approach that only incorporates optical data when it is clear that it will enhance or add value. For example, using a confidence metric to decide when spectral information from Sentinel-2 optical data is effectively enhancing the flood boundary that was derived from the SAR data. The development of functional, multifaceted, and lightweight architecture approaches like LSFUnet Wang et al. (2023) may be an important direction to achieve multi-sensor fusion with efficiency and effectiveness.
5. **Leverage Cloud Platforms for Scalable and Real-Time Processing.**
Cloud platforms such as Google Earth Engine (GEE) were identified as critical enablers of operational deployment in several papers (e.g., Khan et al. (2024); Prakash et al. (2024)). However, there are still computational efficiency and data gap issues that need to be addressed. Future work could continue to remedy computational efficiency of the end-to-end workflows on these platforms; for instance, both automated preprocessing (speckle filtering, terrain correction) for upstream planning and the use of lightweight ML models post-processed itself Wang et al. (2023). Short-term research into model compression, combined with distributed computing, and, at the least, a reasonable and efficient inference pipeline will help enable true real-time flood monitoring, globally.
6. **Advance the Strategic Integration of Polarization and Auxiliary Data.**
The data provides overwhelming evidence of the utility of dual-polarization (VV/VH) SAR data and supplementary layers such as DEMs and HAND index. For example, Wu et al. (2023); Islam et al. (2022) found improved accuracy as a result of their contributions to dual-polar SAR. Future research should focus on systematic experimentation of the roles of different polarization combinations, combinations of each with ancillary data, and to determine optimal combinations in various terrains, rather than simply on ad-hoc additions of data in representing geospatial features of interest. In addition to systematically testing combinations of dual-polar SAR with ancillary data, researchers can develop a standardized approach to resolve resolution mismatch, and other measurement uncertainty, as well as systematically collect and use the spatial context from ancillary data (like DEMs) that provide topographic or hydrologic context to improve distinguishability and label accuracy, and reduce raster false positives in the most complex of landscapes.

In conclusion, this research strategy elucidates and then sets out to resolve the significant, identified gaps, in validation, robustness, generalizability, fusion, scalability, and data integration, illuminated by this review. Addressing these research and priority gaps will enable researchers to accelerate the advancement of ML based flood monitoring from promising prototype to legitimate, operational systems that support process, decision making, and risk in the face of global disaster.

5. Other Information

5.1. Registration and Protocol

5.1.1. Registration

This systematic review was registered with OSF (Registration Link: osf.io/rjxfp).

5.1.2. Protocol Access

The review protocol is available on the OSF database

5.1.3. Protocol Amendments

There was no amendment made after the registration.

5.2. Support

The authors acknowledge the use of Claude [4.5 Sonnet] (developed by Anthropic) for improving the clarity, grammar, and English language style of the final manuscript. The authors confirm that this tool was not used to generate scientific content, perform data analysis, or execute any methodological steps of the systematic review. The authors remain fully responsible for the integrity and conclusions of this publication.

5.3. Competing Interests

The authors declare that they have no competing interests.

5.4. Availability of Data, Code, and Other Materials

All data was uploaded to the GitHub repository.

References

Bai, Y., Wu, W., Yang, Z., Yu, J., Zhao, B., Liu, X., Yang, H., Mas, E., Koshimura, S., 2021. Enhancement of detecting permanent water and temporary water in flood disasters by fusing sentinel-1 and sentinel-2 imagery using deep learning algorithms: Demonstration of sen1floods11 benchmark datasets. Remote Sensing 13, 2220.

Bhadra, T., Chouhan, A., Chutia, D., Bhowmick, A., Raju, P., 2020. Flood detection using multispectral images and sar data, in: International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Springer. pp. 294–303.

De La Cruz, R., Olfindo Jr, N., Felicen, M., Borlongan, N., Difuntorum, J., Marciano Jr, J., 2020. Near-realtime flood detection from multi-temporal sentinel radar images using artificial intelligence. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43, 1663–1670.

Dong, Z., Wang, G., Amankwah, S.O.Y., Wei, X., Hu, Y., Feng, A., 2021. Monitoring the summer flooding in the poyang lake area of china in 2020 based on sentinel-1 data and multiple convolutional neural networks. International Journal of Applied Earth Observation and Geoinformation 102, 102400. URL: https://www.sciencedirect.com/science/article/pii/S0303243421001070, doi:https://doi.org/10.1016/j.jag.2021.102400.

Ebadati, B., Attarzadeh, R., Alikhani, M., Youssefi, F., Pirasteh, S., 2024. Efficient flood detection through hybrid machine learning and metaheuristic methods using sentinel-1. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-3/W3-2024, 35–43. URL: https://isprs-archives.copernicus.org/articles/XLVIII-3-W3-2024/35/2024/, doi:10.5194/isprs-archives-XLVIII-3-W3-2024-35-2024.

El-Haddad, B.A., Youssef, A.M., Pourghasemi, H.R., Pradhan, B., El-Shater, A.H., El-Khashab, M.H., 2021. Flood susceptibility prediction using four machine learning techniques and comparison of their performance at wadi qena basin, egypt. Natural Hazards 105, 83–114.

Fraccaro, P., Stoyanov, N., Gaffoor, Z., La Rosa, L.E.C., Singh, J., Ishikawa, T., Edwards, B., Jones, A., Weldermariam, K., 2022. Deploying an artificial intelligence application to detect flood from sentinel 1 data, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12489–12495.

Ghosh, B., Garg, S., Motagh, M., 2022. Automatic flood detection from sentinel-1 data using deep learning architectures. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3, 201–208.

Ghosh, B., Garg, S., Motagh, M., Martinis, S., 2024. Automatic flood detection from sentinel-1 data using a nested unet model and a nasa benchmark dataset. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science 92, 1–18.

Guo, J., Luan, Y., Li, Z., Liu, X., Li, C., Chang, X., 2021. Mozambique flood (2019) caused by tropical cyclone idai monitored from sentinel-1 and sentinel-2 images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 8761–8772.

Islam, M.S., Sun, X., Wang, Z., Cheng, I., 2022. Fapnet: feature fusion with adaptive patch for flood-water detection and monitoring. Sensors 22, 8245.

Jenifer, A.E., Natarajan, S., 2022. Deepflood: A deep learning based flood detection framework using feature-level fusion of multi-sensor remote sensing images. J. Univers. Comput. Sci. 28, 329–343.

Khamphilung, P., Konyai, S., Slack, D., Chaibandit, K., Prasertsri, N., 2023. Flood event detection and assessment using sentinel-1 sar-c time series and machine learning classifiers impacted on agricultural area, northeastern, thailand. International Journal of Geoinformatics 19.

Khan, N.S., Roy, S.K., Talukdar, S., Billah, M., Iqbal, A., Zzaman, R.U., Chowdhury, A., Mahtab, S.B., Mallick, J., 2024. Empowering real-time flood impact assessment through the integration of machine learning and google earth engine: a comprehensive approach. Environmental Science and Pollution Research 31, 53877–53892.

Kim, J., Kim, H., Jeon, H., Jeong, S.H., Song, J., Vadivel, S.K.P., Kim, D.j., 2021. Synergistic use of geospatial data for water body extraction from sentinel-1 images for operational flood monitoring across southeast asia using deep neural networks. Remote Sensing 13, 4759.

Lam, C.N., Niculescu, S., Bengoufa, S., 2023. Monitoring and mapping floods and floodable areas in the mekong delta (vietnam) using time-series sentinel-1 images, convolutional neural network, multi-layer perceptron, and random forest. Remote Sensing 15, 2001.

Nallapareddy, A., Balakrishnan, B., 2020. Automatic flood detection in multi-temporal sentinel-1 synthetic aperture radar imagery using ann algorithms. International Journal of Computers Communications & Control 15.

Nemni, E., Bullock, J., Belabbes, S., Bromley, L., 2020. Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. Remote Sensing 12, 2532.

Ngo, P.T.T., Hoang, N.D., Pradhan, B., Nguyen, Q.K., Tran, X.T., Nguyen, Q.M., Nguyen, V.N., Samui, P., Tien Bui, D., 2018. A novel hybrid swarm optimized multilayer neural network for spatial prediction of flash floods in tropical areas using sentinel-1 sar imagery and geospatial data. Sensors 18, 3704.

Nguyen, H.D., Van, C.P., Do, A.D., 2023. Application of hybrid model-based deep learning and swarm-based optimizers for flood susceptibility prediction in binh dinh province, vietnam. Earth Science Informatics 16, 1173–1193.

Patil, S., Sawant, S., Joshi, A., 2023. Flood detection using remote sensing and deep learning approaches, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE. pp. 1–6.

Powers, D.M., 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 .

Prakash, A.J., Begam, S., Vilímek, V., Mudi, S., Das, P., 2024. Development of an automated method for flood inundation monitoring, flood hazard, and soil erosion susceptibility assessment using machine learning and ahp–mce techniques. Geoenvironmental Disasters 11, 14.

Saleh, T., Weng, X., Holail, S., Hao, C., Xia, G.S., 2024. Dam-net: Flood detection from sar imagery using differential attention metric-based vision transformers. ISPRS Journal of Photogrammetry and Remote Sensing 212, 440–453.

Sghaier, M.O., Foucher, S., Landry, T., 2019. Multimodal approach for flood monitoring from time-series satellite images combining attribute filters and kohonen map, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 122–125.

Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J., Geertsema, M., Khosravi, K., Amini, A., Bahrami, S., et al., 2020. Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. remote sens 12: 266.

Stateczny, A., Praveena, H.D., Krishnappa, R.H., Chythanya, K.R., Babysarojam, B.B., 2023. Optimized deep learning model for flood detection using satellite images. Remote Sensing 15. URL: https://www.mdpi.com/2072-4292/15/20/5037, doi:10.3390/rs15205037.

Tanim, A.H., McRae, C.B., Tavakol-Davani, H., Goharian, E., 2022. Flood detection in urban areas using satellite imagery and machine learning. Water 14. URL: https://www.mdpi.com/2073-4441/14/7/1140, doi:10.3390/w14071140.

Thapa, A., Horanont, T., Neupane, B., 2022. Parcel-level flood and drought detection for insurance using sentinel-2a, sentinel-1 sar grd and mobile images. Remote Sensing 14, 6095.

Tiampo, K., Woods, C., Huang, L., Sharma, P., Chen, Z., Kar, B., Bausch, D., Simmons, C., Estrada, R., Willis, M., Glasscoe, M., 2021. A machine learning approach to flood depth and extent detection using Sentinel 1A/B Synthetic Aperture radar. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS , 558–561URL: https://doi.org/10.1109/igarss47720.2021.9553601, doi:10.1109/igarss47720.2021.9553601.

Tiampo, K.F., Huang, L., Simmons, C., Woods, C., Glasscoe, M.T., 2022. Detection of flood extent using sentinel-1a/b synthetic aperture radar: an application for hurricane harvey, houston, tx. Remote Sensing 14, 2261.

Tsyganskaya, V., Martinis, S., Marzahn, P., Ludwig, R., 2018. Detection of temporary flooded vegetation using sentinel-1 time series data. Remote sensing 10, 1286.

Wagner, W., Freeman, V., Cao, S., Matgen, P., Chini, M., Salamon, P., McCormick, N., Martinis, S., Bauer-Marschallinger, B., Navacchi, C., Schramm, M., Reimer, C., Briese, C., 2020. Data processing architectures for monitoring floods using sentinel-1. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-3-2020, 641–648. URL: https://isprs-annals.copernicus.org/articles/V-3-2020/641/2020/, doi:10.5194/isprs-annals-V-3-2020-641-2020.

Wang, Z., Wang, X., Li, G., 2023. An extremely lightweight u-net with soft fusion for flood detection using multi-source satellite images, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 6454–6457.

Wedajo, G.K., Lemma, T.D., Fufa, T., Gamba, P., 2024. Integrating satellite images and machine learning for flood prediction and susceptibility mapping for the case of amibara, awash basin, ethiopia. Remote Sensing 16. URL: https://www.mdpi.com/2072-4292/16/12/2163, doi:10.3390/rs16122163.

Wu, H., Song, H., Huang, J., Zhong, H., Zhan, R., Teng, X., Qiu, Z., He, M., Cao, J., 2022. Flood detection in dual-polarization sar images based on multi-scale deeplab model. Remote Sensing 14, 5181.

Wu, X., Zhang, Z., Xiong, S., Zhang, W., Tang, J., Li, Z., An, B., Li, R., 2023. A near-real-time flood detection method based on deep learning and sar images. Remote Sensing 15. URL: https://www.mdpi.com/2072-4292/15/8/2046, doi:10.3390/rs15082046.

Xu, P., Kennedy, G.A., Zhao, F.Y., Zhang, W.J., Van Schyndel, R., 2023. Wearable obstacle avoidance electronic travel aids for blind and visually impaired individuals: A systematic review. IEEE Access 11, 66587–66613. doi:10.1109/ACCESS.2023.3285396.

Yu, H., Luo, Z., Wang, L., Ding, X., Wang, S., 2023a. Improving the accuracy of flood susceptibility prediction by combining machine learning models and the expanded flood inventory data. Remote Sensing 15, 3601.

Yu, H., Wang, R., Li, P., Zhang, P., 2023b. Flood detection in polarimetric sar data using deformable convolutional vision model. Water 15, 4202.

Zhang, C., Wang, R., Chen, J.W., Li, W., Huo, C., Niu, Y., 2023. A multi-branch u-net for water area segmentation with multi-modality remote sensing images, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 5443–5446.

Zhao, Z., Zhang, B., Wu, F., Yang, J., 2023. Deep learning-based flood detection using multi-temporal sentinel-1 sar data: A case study in beijing area, 2023, in: 2023 SAR in Big Data Era (BIGSARDATA), IEEE. pp. 1–4.

Zuhairi, A.H., Yakub, F., Omar, M., Sharifuddin, M., Razak, K.A., Faruq, A., 2024. Imbalanced flood forecast dataset resampling using smote-tomek link, International Exchange and Innovation Conference on Engineering & Sciences.