Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-08

Evaluating machine learning maturity and operational efficacy in database indexing and satellite-based flood mapping: a three-part evidence assessment

Islam, A.S.M Borhanul

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1601

Downloaded from IUB Academic Repository



Evaluating Machine Learning Maturity and Operational Efficacy in Database Indexing and Satellite-Based Flood Mapping: A Three-Part Evidence Assessment

August 2026

Prepared by:

A.S.M Borhanul Islam
ID: 2221128
Sumaiya Tabassum
ID: 2221047
Md Niamal Hafiz Naim
ID: 2220727
Syed Ahanaf Nakibur Rahman
ID: 2131046

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by
Md Asif Bin Khaled
Senior Lecturer
Department of Computer Science and Engineering
Independent University, Bangladesh

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards proper guidelines.

Author Name: A.S.M Borhanul Islam

Signature: .....

Author Name: Sumaiya Tabassum

Signature: .....

Author Name: Md Niamal Hafiz Naim

Signature: .....

Author Name: Syed Ahanaf Nakibur Rahman

Signature: .....

Evaluation Committee

Supervisor
Name:_____ Signature:_____

Internal Examiner 1
Name:_____ Signature:_____

Internal Examiner 2
Name:_____ Signature:_____

Acknowledgement

We would like to sincerely thank our respected sir Md Asif Bin Khaled, Senior Lecturer, Department of Computer Science and Engineering, Independent University, Bangladesh. His dedicated guidance, leadership, perceptive advice, extraordinary patience, insightful suggestions and the generous investment of his valuable time were crucial to the accomplishment of our work. We are really grateful for our supervisor's constant support throughout our academic path.

We also express our sincere gratitude to the Department of Computer Science and Engineering for their outstanding assistance and unwavering support. The creation and implementation of our project was greatly aided by the creative and cooperative environment of the Department. In addition, the support and encouragement have been crucial in shaping our fields of study and research. We are grateful for the resources and opportunities provided to us throughout our investigation, which undoubtedly improved educational experience.

We acknowledge and appreciate the contributions of all who helped us along the way, both directly and indirectly.

Abstract

Machine learning maturity and operational efficacy are evaluated across two unrelated domains, database indexing and satellite-based flood mapping, in a three-part evidence assessment unified not by subject matter but by a shared emphasis on methodological rigor and real-world operational reliability rather than raw reported performance.

Part One is a PRISMA 2020-compliant systematic review of 65 studies (2018–2026) on one-dimensional learned database indexes. Using a novel seven-domain quality framework and GRADE certainty grading, it finds that although learned indexes report a 2–4× median point-lookup speedup over B-trees, certainty of evidence is LOW across all five outcomes, driven by author-led evaluation bias and the 62% of studies that fail to report variance.

Part Two is a PRISMA-ScR scoping review of 73 studies on ML-based flood inundation mapping from Sentinel-1/2 imagery. It documents a rapid shift from CNN encoder–decoders to transformer, diffusion, and self-supervised architectures, and shows multi-sensor fusion outperforming the stronger single-sensor baseline directionally in 65 of 73 sources (37 of 45 sources with a directly comparable numeric value, 82.2%), with gains ranging up to +29.9 percentage points on the primary reported metric, but finds deployment blocked by cloud-dependence, unreleased datasets, and reproducible code released in only a minority of the included studies.

Part Three is an applied benchmark of six SAR flood-detection methods against the Copernicus Global Flood Monitoring service across four monsoon events (2019– 2024) in Roumari Upazila, Bangladesh. The best-performing local method outperforms GFM by 0.17–0.21 Cohen's Kappa, a difference that is statistically significant under paired McNemar tests but that reflects complete, differently built workflows; the comparison does not isolate polarization, thresholding, or masking as the specific cause, and controlled ablation would be needed to make that stronger claim. A crop-calendar overlay further provides seasonal context for interpreting flood exposure, showing that mapped inundation in the 2019 event, though not the largest by extent, overlapped T. Aman transplanting and the Aus harvest; year-specific yield or damage data were not available, so this is reported as phenological exposure rather than an observed agricultural-loss ranking across events.

TABLE OF CONTENTS

Abstract----4
Chapter 1: Project Management----10
1.1 Introduction----10
1.2 Team Structure and Roles----10
1.3 Project Timeline----10
1.4 Project Management Approach----11
Chapter 2: Systematic Literature Review on Learned Indexes----12
2.1 Introduction----12
2.1.1 Background of Machine Learning in Data Systems----12
2.1.2 Problem Statement----12
2.1.3 Research Objective----13
2.2 Literature Review----13
2.2.1 Traditional Database Indexing----13
2.2.2 The Learned Index Revolution----14
2.2.3 Design Variants and Integration Patterns----14
2.2.4 Research Gaps----15
2.3 Methodology----15
2.3.1 PRISMA 2020 Framework----15
2.3.2 Search Strategy----15
Figure 1. Number of Publications by Year----16
Table 1. Database-specific search strings.----16
2.3.3 Seven-Domain Quality Framework----17
Table 2. Seven-domain methodological quality framework for learned-index benchmarking.----18
2.3.4 GRADE Certainty Assessment----19
2.3.5 Data Extraction Process----19
2.3.6 Quality Scoring for Searching SLR----20
2.4 Baseline Models for Learned Indexes----20
2.4.1 B-Trees and LSM-Trees----20
Figure 2. Frequency of baseline comparisons----21
Figure 3. Throughput speedup relative to B-Tree across workload types. Performance degrades as write intensity increases, though learned indexes outperform B-Trees across all categories----22
2.4.2 Adaptive Radix Trees----22
2.5 Results and Analysis----23
2.5.1 Study Selection and Characteristics----23
Figure 4. PRISMA 2020 Flow Diagram (Learned Index SLR).----23
Table 3. Characteristics of Included Studies (Learned Index SLR)----23

2.5.2 Architectural Design Families---- 24
Figure 5. Query Operation Support in Learned Index Designs.---- 25
Figure 6. Update Operation Support Across Learned Index Designs.---- 26
Figure 7. Memory footprint versus lookup latency for learned and classic indexes. Learned indexes push the Pareto frontier toward both smaller memory and lower latency.---- 27
Table 4. Performance Speedup of Learned Indexes vs Baselines---- 27
2.5.3 Dynamic Update Mechanisms----28
Table 5. Four dynamic update mechanism families: representative designs, performance, and failure modes.---- 29
2.5.4 String-Key Handling----29
Figure 8. Distribution of Key Types Evaluated Across Learned Index Studies---- 30
Table 6. String-key encoding strategies: approaches, performance ranges, and failure conditions.---- 30
2.5.5 Concurrency and Production Readiness---- 31
Table 7. Concurrency models across the corpus: designs, thread ceilings, and saturation patterns---- 31
2.5.6 Seven-Domain Quality Scores---- 32
Figure 9. Grouped Percentage Chart Showing the Proportion of Studies with Low Risk, Some Concerns, and High Risk Across Different Bias Domains (Learned Index SLR).----32
Figure 10. Most Frequently Used Datasets in Learned Index Research.---- 33
Table 8. Risk-of-bias summary across seven methodological quality domains (n=65).---33
2.5.7 GRADE Certainty of Evidence----34
Table 9. GRADE certainty summary across five pre-specified outcomes.----35
2.6 Discussion---- 36
2.6.1 Implications For Learned Index Practitioners----36
Figure 11: Multi-threaded throughput scaling of concurrent learned indexes. Lock-free designs (LOFT, Kanva) achieve near-linear scaling while traditional B-Trees hit synchronization bottlenecks.---- 36
2.6.2 Limitations----36
Chapter 3: Scoping Review on Machine Learning for Flood Inundation Mapping----37
3.1 Introduction---- 37
3.1.1 Background---- 37
3.1.2 Problem Statement & Scoping Rationale----39
3.1.3 Aim and Scope Boundaries----39
3.2 Scoping Review Methodology---- 40
3.2.1 PRISMA-ScR Framework and Protocol----40
3.2.2 PCC Framework----40
3.2.3 Eligibility Criteria and Information Sources---- 41
3.2.4 Screening, Data Charting, and Critical Appraisal----41
Figure 12: Growth of Sentinel-1–Sentinel-2 fusion flood-mapping literature, 2016–2026 (n = 73 included studies).----42
3.3 Results---- 44

3.3.1 Selection and Characteristics of Included Studies---- 44
Figure 13. PRISMA Flow Diagram---- 44
Figure 14. Number of Publications by Year---- 45
Figure 15. Fusion-level strategy- pixel-, feature-, decision-, or hybrid-level----46
3.3.2 Landscape of Architectures and Fusion Strategies (RQ1)---- 47
Figure 16: Cloud Handling Methods---- 47
Figure 17. Temporal shift in architecture families, showing transformer and foundation-model approaches emerging only from 2025 onward (RQ1)---- 49
3.3.3 Does Fusion Improve Performance? (RQ2)---- 49
Table 10. Distribution of Reviewed Studies by Model Architecture Family---- 49
Figure 18. Comparison of F1-Score and Overall Accuracy Across Single-Sensor and Sentinel-1–Sentinel-2 Fusion Approaches Grouped by Evaluation Metric---- 51
Table 11. Distribution of Sentinel-1/Sentinel-2 Fusion Studies by Fusion Level and Single-Sensor Baseline Reporting---- 52
Table 12. Benchmark dataset usage across reviewed studies---- 53
Table 13. Distribution of Reviewed Studies (n = 73) by Reporting of Sentinel-2 Cloud-Cover Handling Method and Claimed Operational/Real-Time Deployment---- 54
3.3.4 Validation Rigor and Generalization (RQ3)---- 54
Table 14. Aggregated Sentinel-1/Sentinel-2 fusion performance summary by metric type across reviewed studies---- 55
3.3.5 Operational Readiness, Cloud-Cover Handling, and Deployment (RQ4)---- 56
Table 15. Effect of joint multimodal training on individual-modality branch performance- 56
Table 16. Promising Operational Directions---- 57
3.4 Discussion---- 58
3.4.1 Summary of Evidence---- 58
3.4.2 Conclusions and Implications---- 59
Chapter 4: Flood Mapping Application---- 60
4.1 Introduction---- 60
4.1.1 Problem Statement---- 60
4.1.2 Research Objective---- 61
4.1.3 Relevant Literature Context---- 61
4.2 Methodology---- 63
4.2.1 Study Area---- 63
Figure 19. Study Area Map.---- 63
Table 17. Hydrological Indicator Matrix (FFWC)----63
4.2.2 Data Sources---- 64
Table 18. Sentinel-1 Flood-Window Acquisition Manifest---- 64
4.2.3 Flood Detection Methods---- 65
Figure 20. Overview of the Methodological Framework for SAR-Based Flood Detection, Classification, and Validation Across Four Flood Events.---- 66
Table 20. Method Inventory---- 67

4.2.4 Post-Processing and Validation---67
Table 20. McNemar Comparison Detail for the Best-Performing Local Method Vs. GFM (matched validation points; continuity-corrected asymptotic χ^2 and exact two-sided binomial p)---68
4.2.5 Phenology-Aware Agricultural Exposure Analysis---68
Table 21. Cohen's Kappa (κ) for the best-performing local method vs. GFM, extended (Ext) and restricted (Res) validation windows, across four flood events. Best method is Stage 2 RF (2019 Ext, 2023 Res) or Stage 1 RF (all other cells). Full per-method accuracy metrics (BA, PA, UA, F1) for all six local methods and both windows are intended for the project repository (see Section III.C).--- 69
Table 22. Flood event windows cross-referenced against the regional Kurigram-area crop calendar. Entries describe calendar overlap, not confirmed crop occupancy or damage.- 69
4.2.6 Sentinel-1 Preprocessing--- 69
4.2.7 Training and Calibration--- 70
3.2.8 Validation Framework--- 70
4.2.9 Evaluation Metrics--- 70
4.3 Baseline Models and Comparative Framework---70
4.3.1 Threshold-Based Methods--- 70
4.3.2 Machine Learning Classifiers--- 71
4.3.3 Global Operational Services (GFM)---71
4.4 Results and Analysis---72
4.4.1 Flood Event Character---72
4.4.2 Flood Detection Accuracy--- 72
Figure 21 (a). Sentinel-1 SAR backscatter comparison for 2023 and 2024 showing VV baseline (a), VV flood (b), VH baseline (c), and VH flood (d) scenes over the study area.---72
Figure 21 (b). Sentinel-1 SAR backscatter comparison for 2019 and 2020 showing VV baseline (a), VV flood (b), VH baseline (c), and VH flood (d) scenes over the study area.---73
Figure 22. Balanced Accuracy (%) and F1 Score (%) Comparison of different Methods by Year (2019–2024)---74
4.4.3 Statistical Significance of the Local-GFM Gap---74
Figure 23. Comparison of flood-mapped area (ha) by detection method across agreement categories (Agree, Our-only, and GFM-only)---74
Table 23. Spatial Agreement and GFM Capture Rates--- 75
4.4.4 Spatial Comparison with GFM---75
4.5 Discussion---75
4.5.1 Implications for Flood Monitoring Practitioners--- 75
4.5.2 Limitations---76
Chapter 5: Ethical Considerations--- 77
5.1 Introduction--- 77
5.2 Research and Evidence-Synthesis Integrity--- 77
5.3 Data Provenance, Licensing, and Reproducibility---77
5.4 Humanitarian and Social Implications--- 78
5.5 Environmental and Sustainability Considerations---78

5.6 Scope, Validation, and Responsible Use---------- 78
5.7 Addressing Ethical Issues in Deployment---------- 78
Bibliography---------- 79

Chapter 1: Project Management

1.1 Introduction

This chapter presents the planning and coordination framework used to complete the project. The report combines two literature reviews with an applied flood-detection case study. Chapter 2 presents a systematic literature review of learned database indexes, Chapter 3 presents a scoping review of machine learning for flood inundation mapping, and Chapter 4 presents an applied flood-detection benchmark using satellite data. The project was conducted by four undergraduate researchers from the Department of Computer Science and Engineering, Independent University, Bangladesh, under the supervision of Md Asif Bin Khaled, Senior Lecturer, from August 2024 to September 2026.

1.2 Team Structure and Roles

Name	ID	Role
A.S.M Borhanul Islam	2221128	Member
Sumaiya Tabassum	2221047	Member
Md Niamal Hafiz Naim	2220727	Member
Syed Ahanaf Nakibur Rahman	2131046	Member
Md Asif Bin Khaled	N/A	Supervisor

The project consisted of four student researchers working under a supervisor. The work was divided into three major research workstreams: the learned-index systematic review, the flood-mapping scoping review, and the applied flood-detection case study. All members worked together in each workstream, while also in parallel reviewing the work of each other to identify methodological inconsistencies and improve consistency across the report.

1.3 Project Timeline

The project was conducted from **August 2024 to September 2026**. The three research workstreams were developed partly in parallel, followed by cross-chapter analysis, final compilation, supervisor review, and submission.

Phase	Focus	Duration
1. Project Initiation	Topic selection, scope definition and approval	Aug–Oct 2024

2. Learned-Index SLR Protocol & Methodology Design	PRISMA SLR protocols and search strategies	Nov 2024–Feb 2025
3. Learned-Index SLR	Search, screening, extraction, quality assessment and analysis	Jan–Dec 2025
4. Flood-Mapping Case Study	Data planning, acquisition, modelling and validation	Apr 2025–Apr 2026
5. Flood-Mapping Scoping Review Protocol & Methodology Design	PRISMA ScR protocols, and search strategies	Sep–Oct 2025
6. Flood-Mapping Scoping Review	Screening, extraction and synthesis	Oct 2025–June 2026
7. Cross-Chapter Integration	Results synthesis and ethics analysis	Apr–Jul 2026
8. Finalization	Compilation, supervisor review and revision	Jul–Sep 2026
9. Submission	Final report submission	Sep 2026

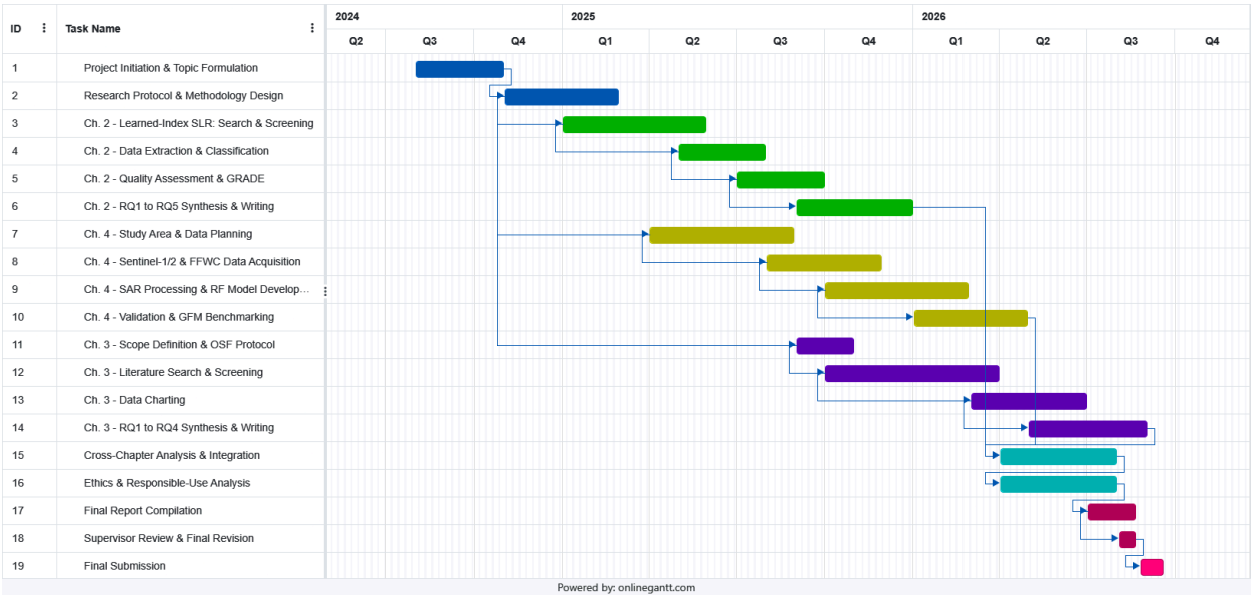


Figure: Gantt Chart on FYDP Report

1.4 Project Management Approach

The project followed a milestone-driven approach rather than fixed development sprints. Major methodological decisions and phase completions were reviewed with the supervisor. Regular team

discussions and shared documentation, including search records, screening decisions and data-provenance information, were used to maintain consistency and traceability throughout the project.

Chapter 2: Systematic Literature Review on Learned Indexes

2.1 Introduction

2.1.1 Background of Machine Learning in Data Systems

Kraska et al. (2018) [1] asked whether a trained machine learning model could replace a traditional index structure. Their core argument was that a model can exploit the statistical distribution of the indexed data to predict a record's position directly, rather than searching a tree, and that this predicts positions with sub-logarithmic lookup cost. On read-heavy workloads, their Recursive Model Index (RMI) [1] reached up to 70% faster lookups than cache-optimized B-trees, alongside an order-of-magnitude reduction in memory footprint, by learning the data's cumulative distribution and predicting record positions from it.

Since that proposal, more than 65 learned index designs have appeared, and most of them target one of three gaps that separated the original prototype from a production system. The first gap was dynamic updates: early learned indexes were static and required full reconstruction after any insertion or deletion. XIndex[2] and FINEdex[3] close this gap with delta buffering; LOFT[4] and Kanva[5] go further, using lock-free CAS-based concurrency to support fully concurrent updates; LIPP[6] instead relies on recursive node splitting, giving $O(\log^2 n)$ amortised insert cost. The second gap was string-key support: the original designs assumed numeric keys, but most real-world databases store variable-length strings. Three encoding strategies address this. Integer projection (SLIPP[7], SLIN[8], SwiftKV[9], DobLIX[10]) maps strings onto integer-like representations. Prefix-based grouping (SIndex[11], RSS[12]) exploits shared prefixes in structured keys such as URLs or file paths. Feature- vector models (SIA[13], SMLP[14]) learn from sub-string features directly. None of the three handles adversarial or high-entropy key distributions well, and SMLP[14] is the clearest failure case in the corpus: a traditional FST baseline outperformed it by 6.0–6.5×. The third gap was system readiness. Production databases need crash consistency, concurrent access, and integration with existing storage engines. APEX[15] and PLIN[16] provide crash consistency for persistent memory, while BOURBON[17] and LearnedKV[18] integrate learned indexes into LSM-tree storage engines.

2.1.2 Problem Statement

Prior work has begun to interrogate the learned-index evidence base empirically. Liu et al. [19] and Sun et al. [20] benchmark multiple designs under shared experimental harnesses; Wongkham et al. [21] independently stress-tests update mechanisms across designs rather than relying on proposing-authors' own claims; and Maltry and Dittrich [22] offer a critical analysis of RMI-style designs. These efforts establish that cross-design, non-author benchmarking is possible and valuable. However, none applies a systematic, PRISMA-compliant methodology with formal per-domain risk-of-bias scoring and outcome-level certainty grading across the full landscape of 65+ published designs. The field has experienced rapid publication pace (65 studies in 8 years), with each new design claiming improvements

over previous work, but no synthesis has yet quantified how much confidence the aggregate evidence base actually warrants, outcome by outcome. Critical questions remain open at the level of the field rather than the individual paper: are baseline comparisons conducted fairly across the corpus as a whole? Is variance quantification the exception or the norm? Do evaluations reflect production-representative workloads systematically, or only in isolated cases? Without a structured answer to these questions, database engineers considering learned-index adoption have individual benchmarking papers to draw on, but no field-level certainty assessment to calibrate how much weight those papers should carry. The overarching question is how far 1D ordered learned indexes have progressed beyond early prototypes and how credible the evidence for reported improvements is.

RQ1 (Architectural Design Families): What design families exist for 1D ordered learned indexes, and what distinguishes their architectures?

RQ2 (Dynamic Update Mechanisms): What update mechanism families are used, and what algorithmic trade-offs do they impose across workload and hardware regimes?

RQ3 (String-Key Handling): How do current designs represent and index variable-length string keys, and where do those strategies break down?

RQ4 (Concurrency, Persistence, and Deployment Readiness): Which systems support concurrent access and crash recovery, and how credible is the performance evidence?

RQ5 (LSM-Tree and Key-Value Store Integration): How have learned indexes been embedded in LSM-based storage engines and key-value stores, and which approaches work under realistic workloads?

2.1.3 Research Objective

To systematically review one-dimensional ordered learned indexes, constructing taxonomies of design variants, update mechanisms, string-key strategies, and system-integration patterns, then assessing evidence quality through a seven-domain framework with GRADE certainty grading across 65 published studies spanning 2018-2026.

2.2 Literature Review

2.2.1 Traditional Database Indexing

Database indexes exist to find records quickly. For decades, this task has been handled by B-trees and hash tables, structures that have done the job reliably. However, neither structure pays attention to the data it stores, they treat all key distributions the same way, regardless of the underlying data's shape or patterns. This limitation set the stage for a fundamentally different approach: in 2018, Kraska et al. [1] proposed replacing the index with an ML model that learns the data's distribution and predicts where a record sits, rather than relying on a fixed, data-agnostic structure. Their RMI (Recursive Model Index) [1] stacked small models in a hierarchy, each narrowing the search until reaching the target position, achieving up to 70% faster lookups than a cache-optimized B-tree with an order-of-magnitude less memory on read-heavy workloads.

2.2.2 The Learned Index Revolution

In 2018, Kraska et al.[1] proposed replacing the index with an ML model that learns the data's distribution and predicts where a record sits, rather than relying on a fixed, data-agnostic structure like a B-tree or hash table. Their RMI (Recursive Model Index)[1] stacked small models in a hierarchy, each narrowing the search until reaching the target position, achieving up to three times the speed of a B-tree with less memory on read-heavy workloads. Building on that foundation, the PGM-index[23] provided provable worst-case bounds on space and query time, the FITing-Tree[24] let users set a maximum prediction error, and ALEX [25] became the first design built for frequent insertion. From there, the field closed three gaps separating early prototypes from production use: dynamic update support (LIPP[6], XIndex[2], FINEdex[3], with LOFT[4] and Kanva[5] extending this to fully lock-free operation), string-key support (SIndex[11], RSS[12], SLIPP[7], LITS[26], though adversarial distributions remain unsolved) and system readiness (APEX[15] and PLIN[16] for persistent-memory crash-consistency, BOURBON[17] and LearnedKV[18] for LSM-tree integration). The strongest industrial evidence to date is Google Bigtable, where Abu-Libdeh et al.[27] reported a 36% drop in mean point-lookup latency.

2.2.3 Design Variants and Integration Patterns

Since 2018, over 65 distinct learned index designs have been proposed, addressing three gaps that separated early prototypes from production systems. The dynamic update gap was addressed by LIPP [6], classified as a tree-structure hybrid design that embeds learned models inside modified trees, achieving $O(\log^2 n)$ amortised insert cost. XIndex [2] employed optimistic concurrency control for scalable concurrent reads during updates. FINEdex[3] used a delta-buffering/level-bin mechanism for dynamic updates, paired with fine- grained locking for concurrency control. LOFT[4] and Kanva[5] extended these mechanisms to fully lock-free operation for dynamic workloads.

The string-key gap was addressed through diverse encoding strategies. SIndex [11] used a prefix-based grouping approach, achieving an 89% throughput gain over XIndex under a 9:1 read-write workload on structured keys with shared prefixes such as URLs and paths. RSS[12] used the same prefix-based grouping strategy, producing an index 7×–70× smaller than ART/HOT on similarly structured key sets. SLIPP[7] combined recursive node splitting with integer projection techniques. LITS[26] reached up to 5.31× the insertion performance of SIndex[11].

The system readiness gap was addressed by designs targeting specific deployment environments. APEX[15] relies on logical undo-redo logging for crash consistency on persistent memory, while PLIN[16] uses redo logs with atomic upserts. BOURBON[17] replaced the SSTable block index in WiscKey[17] with a learned model, marking the first LSM-tree learned-index integration. LearnedKV[18] extended this with a two-layer architecture, using the LSM tree for writes and a learned index for GC-stable data. The strongest industrial system-integration evidence came from Google Bigtable, where Abu-Libdeh et al.[27] reported a 36% drop in mean point-lookup latency under controlled experimental evaluation on Bigtable's serving stack, the strongest system-integration evidence in the corpus.

2.2.4 Research Gaps

Existing cross-design efforts [19–22] demonstrate that independent benchmarking of learned indexes is both feasible and revealing, but each addresses a specific slice of the problem, a shared benchmark harness, an update-mechanism stress test, or a critique of one design family, rather than a systematic, risk-of- bias-scored synthesis spanning the full corpus. No prior work has combined PRISMA 2020 methodology with a domain-specific quality framework and GRADE-style certainty grading to assess the learned-index literature as a whole. The field also lacks standardized evaluation protocols for adversarial robustness, crash consistency, distribution drift, and concurrent access, and the predominance of author-led evaluations within the proposal literature itself creates a structural positive-result bias that a handful of independent benchmarking papers has not yet been able to fully counteract.

2.3 Methodology

2.3.1 PRISMA 2020 Framework

PRISMA 2020 guidelines[28] adapted for the computer systems domain where conference publications carry equal or greater weight than journal articles. The adaptation reflects the publication culture of computer systems research, where top-tier conferences (SIGMOD, VLDB, OSDI, SOSP) serve as primary venues for novel contributions, and journal publications often extend conference papers with additional evaluation. This review's protocol was registered retrospectively, after search execution and initial data collection had already begun, and the review's analytical emphasis shifted during the course of the work relative to that initial protocol; this deviation is disclosed here rather than presented as prospective registration, and is treated as a limitation (Section 1.6.2).

2.3.2 Search Strategy

Seven information sources were searched on 10 January 2026: five bibliographic databases (ACM Digital Library, IEEE Xplore, ScienceDirect, MDPI, SpringerLink), one preprint repository (arXiv), and one search engine (Google Scholar). Search strings paired core terms (learned index, recursive model index, model-based index) with modifiers targeting dynamic behaviour and deployment context, for example: ("Learned Index*" OR "recursive model index*") AND (Dynamic OR Update* OR lookup*) AND ("String Key*" OR "Variable-length" OR "Persistent Memory" OR Concurrent*). Reference lists were not hand-searched, and grey literature was limited to arXiv. Studies were eligible if published between 1 January 2018 and 10 January 2026, the period beginning with the year Kraska et al. introduced the RMI[1]. Full source-specific queries, run dates, result counts, and deduplication rules are archived in the project GitHub repository. The string reported here is the conceptual query; platform syntax (field tags, wildcards, and filters) follows each source’s native interface and is recorded in that archive so a third party can re-run each search.

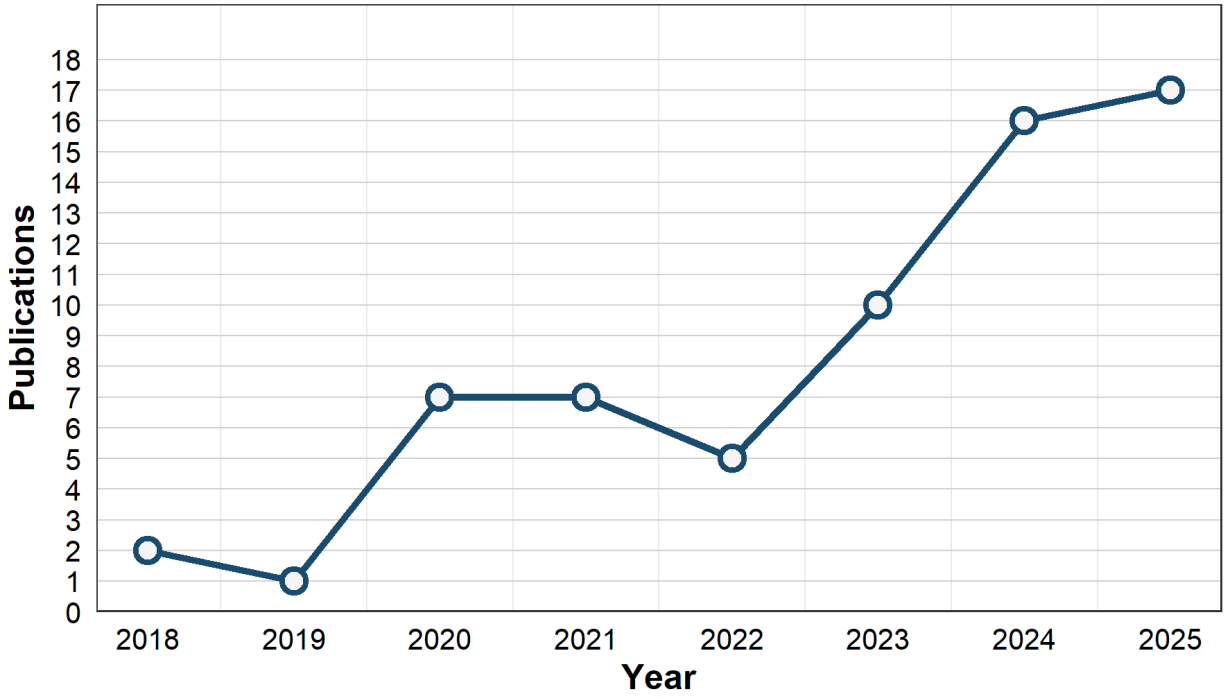


Figure 1. Number of Publications by Year

Table 1. Database-specific search strings.

Database	Search Date	Search String (Abbreviated)	Filters
IEEE Xplore	10/01/2026	("Abstract":"Learned Index*" OR "recursive model index*") AND ("Full Text":Dynamic OR Update* OR lookup* OR "String Key*")	2018–2026
ScienceDirect	10/01/2026	"Learned Index" OR "recursive model index"	2018–2026
SpringerLink	10/01/2026	("Learned Index*" OR "recursive model index*") AND (Dynamic OR Update* OR lookup*) AND ("String Key*" OR "Variable-length" OR "Persistent Memory" OR Concurrent*)	2018–2026
MDPI	10/01/2026	"Learned Index" AND (Dynamic OR Update OR lookup OR "String Key")	2018–2026
ACM	10/01/2026	[All:"learned index*"] AND [All:dynamic OR update* OR lookup*] AND [All:"string key*" OR "variable-length" OR "persistent memory" OR concurrent*]	2018–2026
arXiv	10/01/2026	"Learned Index" AND (Dynamic OR Update OR lookup OR "String Key")	2018–2026

Google Scholar	10/01/2026	("Learned Index*" OR "recursive model index*") AND (Dynamic OR Update* OR lookup*) AND ("String Key*" OR "Variable-length" OR "Persistent Memory" OR Concurrent*)	2018–2026

2.3.3 Seven-Domain Quality Framework

A novel seven-domain methodological quality framework, drawing on Kitchenham and Charters' (2007)[29] SLR methodology, was developed specifically for assessing benchmarking evidence quality in learned-index studies. Each domain comprises five questions scored 0 (absent), 1 (partial), or 2 (adequate), for a maximum of 10 points per domain.

Domain D1 (Baseline Fairness & Experimental Control) assessed experimental control and baseline parity, evaluating whether baseline methods were clearly described and fairly implemented, whether experiments were conducted on the same hardware with controlled resources, whether confounding factors were adequately controlled or reported, whether dataset characteristics were adequately described, and whether hyperparameters were optimized fairly across all methods.

Domain D2 (Bias in Selection of Datasets/Workloads) assessed distribution diversity and workload coverage, evaluating whether dataset selection criteria were clearly described, whether datasets covered diverse distributions, whether diverse string characteristics were tested, whether both synthetic and real-world datasets were included, and whether datasets were selected without evidence of cherry-picking.

Domain D3 (Bias in Implementation of Methods) assessed code availability and reproducibility, evaluating whether the learned index implementation was clearly described, whether source code was publicly available or reproducible, whether baseline implementations were from reputable sources, whether model architectures and training procedures were described, and whether the implementation matched the intended design.

Domain D4 (Bias in Experimental Execution) assessed failure coverage and concurrency reporting, evaluating whether reported experiments matched the described methodology, whether insert/update/delete patterns were clearly specified, whether concurrency levels and thread interactions were described, whether failure scenarios were realistic and recovery was tested, and whether workload parameters were clearly specified.

Domain D5 (Bias Due to Incomplete Reporting) assessed metric coverage and limitation disclosure, evaluating whether all relevant metrics were reported, whether negative results or failure cases were discussed, whether limitations were clearly acknowledged, whether results were reported for all tested configurations, and whether outliers or unexpected behaviours were explained.

Domain D6 (Bias in Performance Measurement) assessed variance, warm-up, and confidence-interval reporting, evaluating whether performance measurement methodologies were clearly described, whether multiple runs were performed to account for variability, whether confidence intervals or error bars were reported, whether warm-up periods were used to eliminate cold-start effects, and whether measurement overheads were accounted for.

Domain D7 (Bias in Selective Reporting) assessed result completeness and reproducibility, evaluating whether results were reported for all datasets mentioned, whether both favourable and unfavourable

comparisons were shown, whether results were reported without evidence of p-hacking or selective reporting, whether reproducibility materials were publicly available, and whether all promised experiments were reported.

Domains were classified by risk of bias: Low (8–10), Moderate (5–7), Serious (2– 4), or Critical (0–1).

Table 2. Seven-domain methodological quality framework for learned-index benchmarking.

Domain	**Focus**	**Assessment Questions (scored 0–2 each; max 10)**
D1	Baseline Fairness & Experimental Control	Q1: Were baseline methods clearly described and fairly implemented? Q2: Were experiments conducted on the same hardware with controlled resources? Q3: Were confounding factors adequately controlled or reported? Q4: Were dataset characteristics adequately described? Q5: Were hyperparameters optimized fairly across all methods?
D2	Bias in selection of datasets/worklo ads	Q1: Were dataset selection criteria clearly described? Q2: Do datasets cover diverse distributions? Q3: Were diverse string characteristics tested? Q5: Were datasets selected without evidence of cherry-picking? Q5: Were datasets selected without evidence of cherry-picking?
D3	Bias in implementation of methods	Q1: Is the learned index implementation clearly described? Q2: Is source code publicly available or reproducible? Q3: Are baseline implementations from reputable sources? Q4: Were model architectures and training procedures described? Q5: Does implementation match intended design?
D4	Bias in experimental execution	Q1: Do reported experiments match methodology described? Q2: Are insert/update/delete patterns clearly specified? Q3: Are concurrency levels and thread interactions described? Q4: Are failure scenarios realistic and recovery tested? Q5: Were workload parameters clearly specified?
D5	Bias due to incomplete reporting	Q1: Are all relevant metrics reported? Q2: Are negative results or failure cases discussed? Q3: Are limitations clearly acknowledged? Q4: Are results reported for all tested configurations? Q5: Are outliers or unexpected behaviours explained?
D6	Bias in performance measurement	Q1: Are performance measurement methodologies clearly described? Q2: Were multiple runs performed to account for variability? Q3: Are confidence intervals or error bars reported? Q4: Were warm-up periods used to eliminate cold-start effects? Q5: Are measurement

Domain	Focus	Assessment Questions (scored 0–2 each; max 10)
		overheads accounted for?
D7	Bias in selective reporting	Q1: Are results reported for all datasets mentioned? Q2: Are both favorable and unfavorable comparisons shown? Q3: Were results reported without evidence of p-hacking or selective reporting? Q4: Are reproducibility materials publicly available? Q5: Were all promised experiments reported?

2.3.4 GRADE Certainty Assessment

Evidence certainty was graded using a GRADE-informed framework [30] applied to five pre-specified outcomes: point-lookup performance, space/memory efficiency, range-query performance, update performance, and concurrency scaling. We do not claim formal GRADE compliance as defined for clinical evidence profiles. Instead, we use GRADE's certainty labels (High, Moderate, Low, Very Low) and its core logic of starting from study design and then downgrading for limitations and publication bias. Because every included study was proposed and evaluated by the same team, all five outcomes were rated Low certainty after downgrading for study limitations (self-evaluation) and publication bias.

All five pre-specified outcomes received a LOW certainty rating. The LOW rating reflects not the volume of evidence but its origin: across all 65 studies, the team proposing each method was also the team that evaluated it, an inherent consequence of the inclusion criteria requiring novel index proposals, not that a single research team authored all 65 studies. This creates structural positive-result bias, not because the papers are dishonest, but because a field where every performance claim comes from its own proposing authors, with no external replication study inside the corpus, cannot yet be considered settled evidence. For each outcome, certainty was downgraded for study limitations (self- evaluation) and for publication bias.

2.3.5 Data Extraction Process

Data were extracted into a structured spreadsheet across bibliographic, structural, and five outcome-specific fields. Bibliographic fields captured authors, year, title, venue, and peer-review status. Structural fields capture key type (integer, float, string), update support (insert, delete, update, bulk load), code availability, and hardware deployment regime.

Studies were tiered by their primary methodological contribution into four deployment-stage tiers: Tier 1 (Novel Design) for papers proposing fundamentally new architectures; Tier 2 (System Integration) for papers integrating learned indexes with existing systems; Tier 3 (Foundational Proposal) for theoretical or proof-of-concept contributions; and Tier 4 (Industrial Deployment) for papers reporting production system results. Studies were further grouped into six hardware-deployment regimes: CPU in-memory numeric, string-key specialised, NVM/persistent memory, GPU/hardware-accelerated, DPU/RDMA/networked, and LSM-tree/key-value store integration.

2.3.6 Quality Scoring for Searching SLR

Each of the seven quality domains (D1–D7) was scored using the five-question rubric, with each question rated 0 (absent), 1 (partial), or 2 (adequate), giving a maximum of 10 points per domain. Domain scores were aggregated into risk-of-bias ratings (Low, Moderate, Serious, or Critical) reported per domain across the 65 studies, as well as an overall per-study rating. No formal inter- rater reliability statistic such as Cohen's Kappa was calculated for this review; the paper notes the absence of such a measure (for the eligibility- screening process) as an acknowledged limitation, not something that was assessed.

The quality scoring revealed several systematic patterns: Domain D6 (Performance Measurement Rigour) was the dominant weakness, with only 5% of studies rated Low risk (54% Serious, 8% Critical). Domains D2 (Dataset & Workload Realism, 58% Low) and D4 (Experimental Execution, 57% Low) were the next weakest. By contrast, Domain D5 (Reporting Completeness) scored highest at 95% Low risk, followed by D7 (Selective Reporting) at 92%, with D1 (Baseline Fairness) and D3 (Implementation Transparency) both at 89%. This pattern suggests the field describes its setups, compares against recognised baselines, and reports what it set out to measure reasonably well, but lags badly in variance reporting, confidence intervals, and warm-up procedures, and shows weaker performance on dataset diversity and experimental execution reporting.

2.4 Baseline Models for Learned Indexes

2.4.1 B-Trees and LSM-Trees

B-Tree is the dominant baseline in learned-index evaluations, used for comparison in 80% of the reviewed papers, while 12% of studies include no classic baseline comparison at all. Across workload types, learned indexes consistently outperform B-Trees in throughput, though the margin narrows as write intensity increases, since write-heavy conditions strain the update mechanisms most learned designs bolt on around a fundamentally read-optimized model. This pattern recurs in individual mechanism families as well: the gapped- array-reservation family (ALEX[25] , LIFT[31] , LOFT[4] , COLIN[32] , NFL[33] , ShapeShifter[34] , SWIX[36] , FITing-Tree[24] , CSV[37]) reports speedups of 1.06×–14× over B-tree with O(log m) amortised update cost, but space overhead grows with gap density and write bursts trigger retraining stalls that erode the throughput advantage.

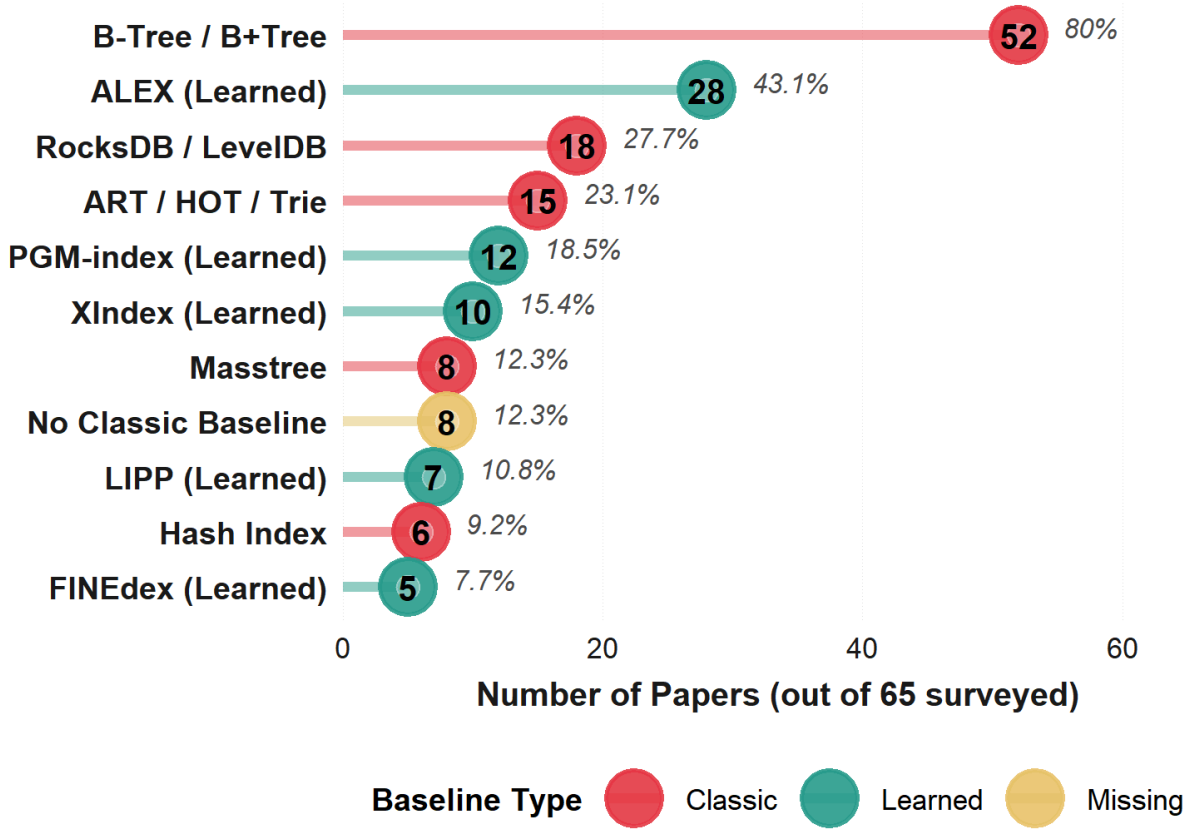


Figure 2. Frequency of baseline comparisons

Fifteen studies demonstrate LSM-tree or key-value store integration, and the primary integration point across this line of work is the SSTable block index. BOURBON[17] was among the first to pursue this integration, followed by TieredKV[38] , LearnedKV[18] , SwiftKV[9] , LeaderKV[39] , HL-LSM[40] and LCKV[41] . BOURBON[17] replaces the WiscKey[17] block index with a learned model, yielding 1.23×–1.78× point-lookup gains, though the benefit dilutes once range queries exceed roughly 500 keys. LearnedKV[18] takes a two-layer approach LSM for writes, a learned structure for GC-stabilised data reporting 4.32× read and 1.43× write improvement over RocksDB+[18] along with a 2.98× reduction in index size, with the largest gains on large datasets at moderate write ratios; TieredKV[38] achieves similar read/write gains through a different tiering architecture. LeaderKV[39] adds hotness-aware compaction to a learned read path, reaching 1.75×–3.68× over RocksDB/LevelDB and 1.75× over BOURBON[17] on skewed, read-heavy access patterns, while HL-LSM's[39] read-hotness-aware design delivers a more modest 22%–48% throughput improvement over WiscKey[17]/LevelDB using a metadata-only, bulk-load-only learned layer. Across this regime as a whole, reported gains range from 1.23×–4.32× over RocksDB/WiscKey baselines, with index-size reductions from near-zero up to 2.98×, but the gains consistently diminish under write-heavy conditions and long range scans, reflecting the underlying LSM compaction lifecycle. BOURBON, LearnedKV, and TieredKV are described as the most production-ready options in this space, though the certainty of evidence for this outcome remains rated LOW.

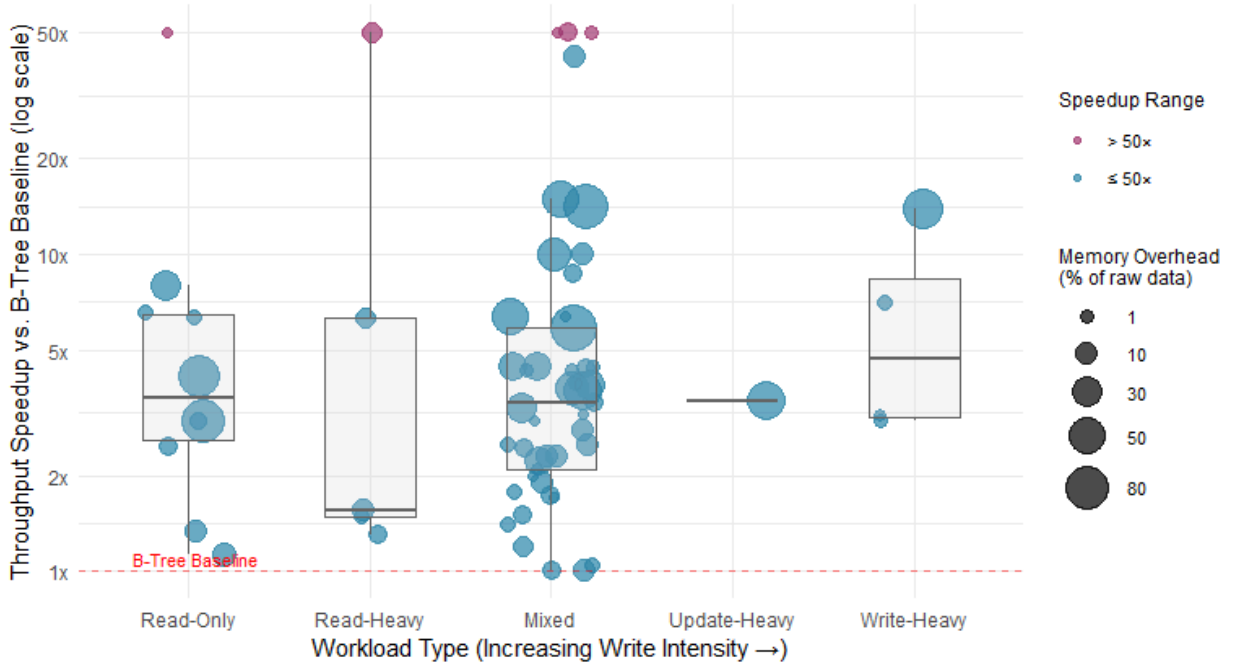


Figure 3. Throughput speedup relative to B-Tree across workload types. Performance degrades as write intensity increases, though learned indexes outperform B-Trees across all categories

2.4.2 Adaptive Radix Trees

The Adaptive Radix Tree (ART) provides a high-performance comparison baseline for string-key learned indexes. ART efficiently handles diverse key distributions without wasting memory on sparse nodes, achieving high throughput through cache-friendly memory layouts, making it a demanding baseline for string key learned indexes to exceed. Several string-key learned indexes report comparison against ART, though the fairness of these comparisons varies across studies.

Key distribution strongly influences relative performance: on uniform random strings, learned indexes can match or exceed ART throughput, while on skewed real-world distributions (URLs, file paths, natural-language text), ART frequently outperforms learned alternatives due to its adaptive node sizing that naturally handles variable-length keys and common prefixes.

2.5 Results and Analysis

2.5.1 Study Selection and Characteristics

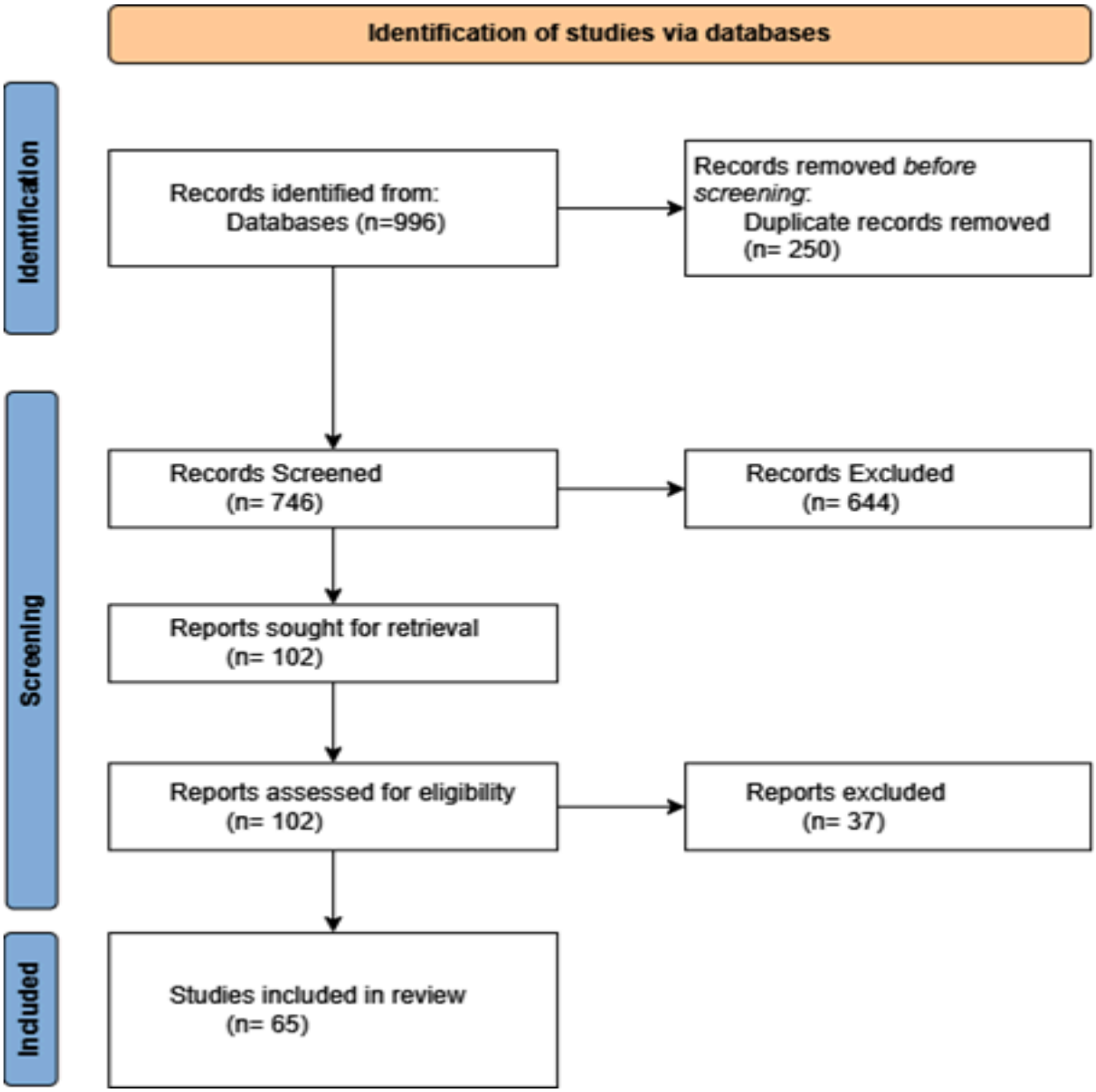


Figure 4. PRISMA 2020 Flow Diagram (Learned Index SLR).

Table 3. Characteristics of Included Studies (Learned Index SLR)

Category	Breakdown

Screening funnel	746 screened → 102 full-text assessed → 65 included after full-text review [1–18, 23–27, 32–73]
Publication window	Jan 2018 – Jan 2026
Publication venue	Conference publications dominated with the remainder in journals and preprint servers
Hardware-deployment regime	A: CPU in-memory numeric (30); B: string-key specialised (10); C: NVM/persistent memory (6); D: GPU/hardware-accelerated (4); E: DPU/RDMA/networked (2); F: LSM-tree/key-value integration (15)

Note: hardware-regime counts sum to 67 rather than 65 because two studies carry secondary regime tags.

2.5.2 Architectural Design Families

Analysis of the 65 studies identified six primary architectural families that characterise the design space of learned indexes, each reflecting a distinct strategy for approximating key distributions.

RMI-Style Hierarchical Models. The foundational approach organises a multi-level hierarchy of stacked regression models, where each level narrows the search space for the next. Later work builds on this foundation by adding adaptive node splitting, delta buffering for write support, and lock-free CAS-based concurrency mechanisms.

PGM-Style Piecewise Geometric Models. This approach takes a more formally grounded route, constructing a recursive structure of piecewise linear segments with a guaranteed maximum error ε. This allows performance bounds to be proven analytically rather than measured empirically. Several later designs extend this line of work.

FITing-Tree-Style Data-Aware Structures. This family allows the underlying data distribution to directly determine segment boundaries. This makes the approach highly effective on smooth distributions, which compress well, but exposes its limits on skewed or adversarial data. A number of subsequent designs build on this data-aware philosophy.

Tree-Structure Hybrid Learned Indexes. This family embeds learned models inside modified tree structures specifically to keep updates manageable, with representative designs achieving amortised insert costs on the order of $O(\log^2 n)$. Several other designs each offer variations on this hybrid strategy.

LSM-Tree Integrated Learned Indexes. Fifteen of the 65 studies integrate learned indexes into LSM-tree or key-value store systems, making this one of the largest families. Early work in this area has since been extended by numerous designs targeting different points in the read/write tradeoff.

Hardware-Specialised Learned Indexes. Six studies treat the target hardware as a first-class design constraint rather than an afterthought. Several designs are built specifically around persistent memory semantics, while others target GPU acceleration or offload index operations to DPUs.

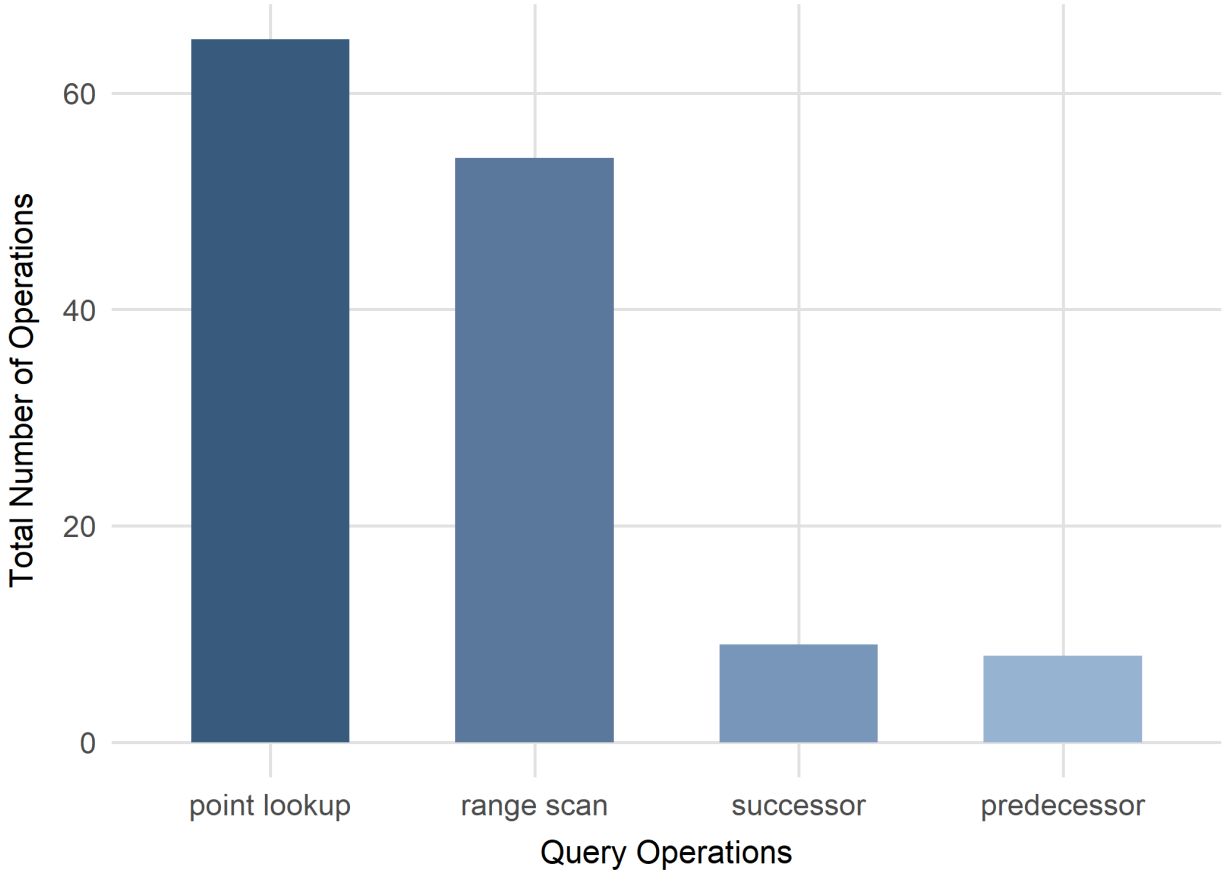


Figure 5. Query Operation Support in Learned Index Designs.

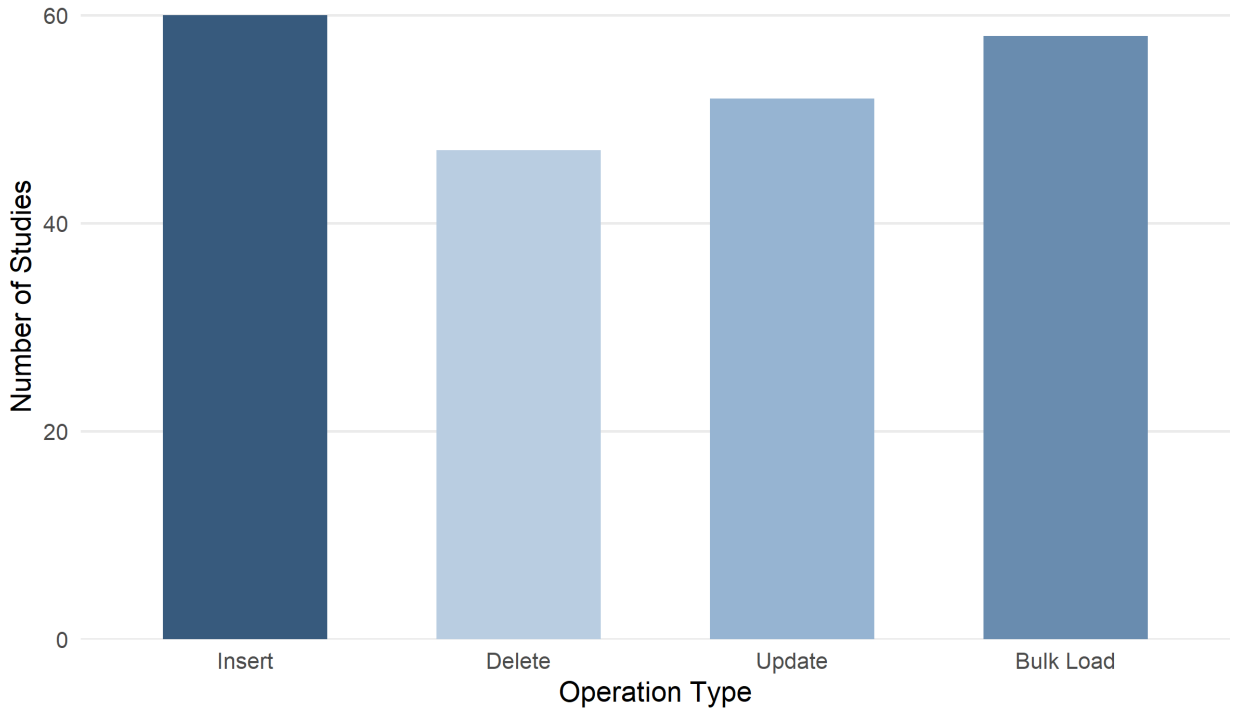


Figure 6. Update Operation Support Across Learned Index Designs.

These two figures summarise, for each of the six architectural families above, which query types (point lookup, range scan) and which update operations (insert, delete, bulk load) are supported.

Together, these six families illustrate a clear progression in the field: from purely statistical hierarchical models, through formally bounded and data-driven geometric approaches, to hybrid tree structures and system-level integrations that address the practical demands of dynamic updates, and finally to hardware-specialised designs that trade portability for platform-specific performance gains.

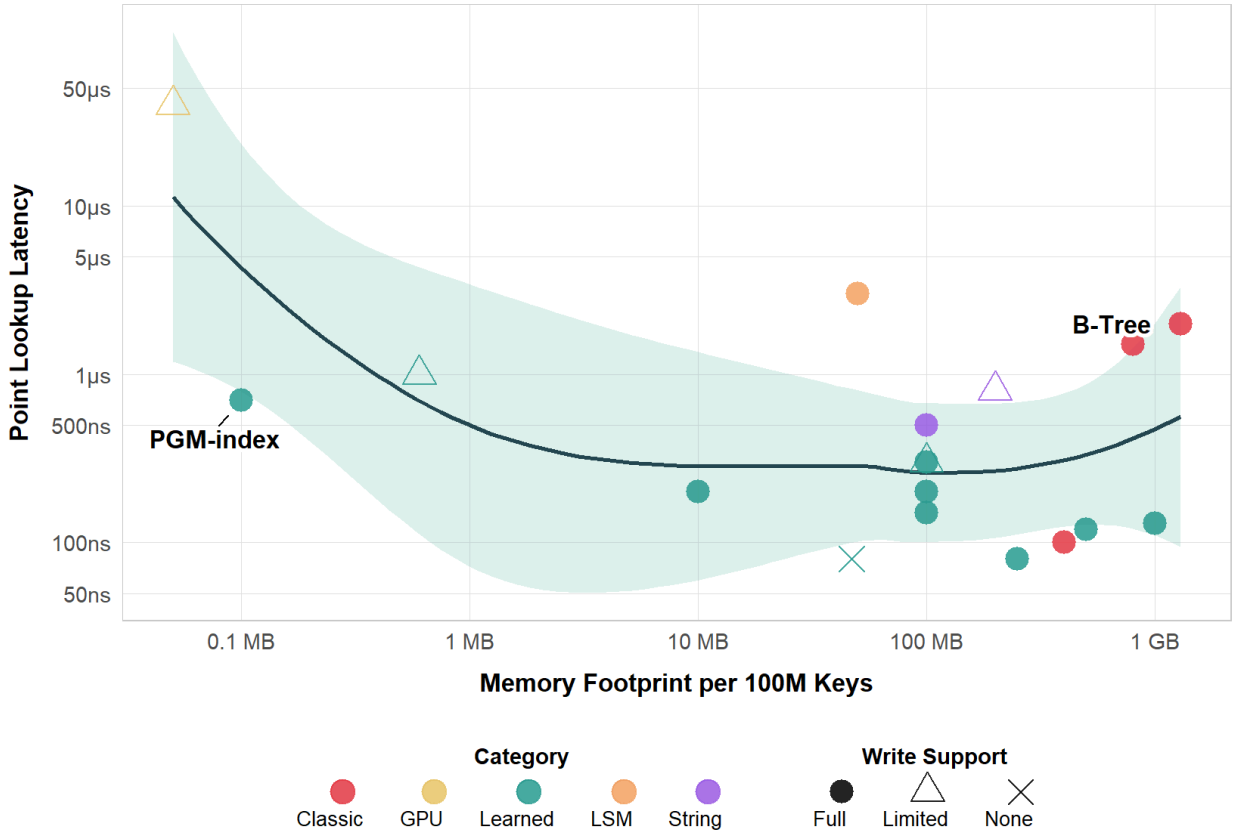


Figure 7. Memory footprint versus lookup latency for learned and classic indexes. Learned indexes push the Pareto frontier toward both smaller memory and lower latency.

Table 4. Performance Speedup of Learned Indexes vs Baselines

Design	Baseline Compared	Speedup / Improvement
BOURBON[17]	WiscKey (SSTable block index)	1.23×–1.78× point lookup
LearnedKV[18]	RocksDB	4.32× read; 1.43× write
TieredKV[38]	LSM baseline	4.32× read; 1.43× write
SwiftKV[9]	RocksDB / WiredTiger	1.5×–2.1× (RocksDB); 1.2×–3.2× (WiredTiger)
LeaderKV[39]	RocksDB / LevelDB / BOURBON	1.75×–3.68× (RocksDB/LevelDB); 1.75× (BOURBON)
HL-LSM[40]	WiscKey / LevelDB	22%–48% throughput increase

XStore[67]	Traditional RDMA KV store	5.9× read speedup
Bigtable Learned Index[42]	Production baseline	36% drop in mean point-lookup latency
LCKV[41]	Mixed baseline	4.73% faster (read-only); on par (mixed)
LearnedStore[53]	LeanStore B-tree	2.29× read speedup; 6.84× tail latency reduction
SLR Search[54]	Binary search (LevelDB)	12.7% improvement
AStore[63]	XStore	Up to 2×
XIndex[2]	Masstree	3.2×
XIndex-R[56]	Masstree	3.2×–4.4×
FineStore[57]	Competing concurrent/distributed KV systems	1.8×–2.5×
APEX (NVM)[15]	Baseline NVM index	15×–16.38× insert speedup
DALdex[66]	APEX / PLIN	1.07×–6.34×
LITS[26]	SIndex	Up to 5.31× insertion throughput
SIndex[11]	XIndex (9:1 read-write)	89% throughput gain
SIA[13]	YCSB baseline	2.6× avg throughput
SMLP (negative result)[26]	FST	FST outperforms by 6.0×–6.5×

2.5.3 Dynamic Update Mechanisms

Of the 65 studies, 55 handle inserts, 44 handle deletes, and 43 support in-place updates. Ten papers are bulk-load-only: RSS[12] , SMLP[14] , RadixSpline[60] , SLR-LevelDB[64] , RMI[1] , HL-LSM[40] , LearnedKV[18] , Bigtable Index[27] , SLIN[8] , and TieredKV[38] . This split highlights a critical gap in the field: the majority of learned index designs cannot accommodate changing data without complete reconstruction, limiting their applicability to read-heavy archival workloads.

Update mechanisms were classified into four categories. Append-Only Structures accommodate new keys by extending the model without restructuring existing predictions, achieving strong insert throughput but degrading space efficiency under sustained writes. Adaptive Rebuilding periodically restructures portions of the index based on configurable thresholds, balancing write amplification against read performance degradation. Copy-on-Write creates new versions of modified regions, enabling lock-free concurrent reads during updates. LSM-Tree Integration delegates dynamic management to the LSM-tree's compaction process, replacing only the SSTable block index with a learned predictor.

Table 5. Four dynamic update mechanism families: representative designs, performance, and failure modes.

Mechanism Family	Representative Designs	Insert/Update Speedup	Concurrency Ceiling	Primary Failure Mode
Gapped Array Reservation	ALEX [25], LIFT [31], LOFT [4], COLIN [32], NFL [33], ShapeShifter [34], SWIX [36], FITing-Tree [24], CSV [37]	1.06×-14× over B-tree; O(log m) amortised cost	Up to 80 threads (LOFT [4]); 64 threads (SALI [51])	Space overhead grows with gap density; write bursts trigger retraining stalls that degrade throughput.
Delta Buffering / Level-Bin	XIndex [2], FINEdex [3], DobLIX [10], XIndex-R [68], SIndex [11], EWALI [59], DTtree [72]	1.6×-3.2× insert improvement; 27 µs retraining latency per level-bin	24 threads (FINEdex [3]); 128 threads (DyTIS [71])	Retraining lag (~300 µs per ~1,400 keys) under write bursts
Recursive Node Splitting	LIPP [6], SLIPP [7], DILI [56], Chameleon [52], Sig2Model [57]	$O(\log^2 n)$ amortised insert; 2.8×-2.9× lookup vs ALEX [25]	Up to 128 threads (Sig2Model [57] peaks at 7)	Cascading splits under adversarial key sequences; memory overhead
Hardware-Specific Mechanisms	APEX [15], PLIN [16], WIPE [69], DALdex [66], PFLX [54]	15×-16.38× insert speedup (APEX [15]); 1.07×-6.34× vs APEX/PLIN (DALdex [66])	Platform-dependent; DALdex [66] offloads to DPU	NVM ordering semantics; five incompatible crash-recovery strategies

2.5.4 String-Key Handling

Only 20 of 65 (30%) studies evaluate variable-length string keys. Three distinct encoding strategies emerged, summarised with representative designs in Table 6 below: Integer Projection maps string keys onto a numeric surrogate so existing numeric learned-index techniques apply directly, performing well on short, fixed-length, uniformly distributed keys but breaking down on long, high-entropy, or adversarial inputs; Prefix-Based Grouping exploits shared key prefixes (as in URLs or file paths) for compact routing, but degrades as prefix diversity increases and tree depth grows unbounded; and Feature-Vector Models extract rich sub-string features for the underlying model, working well on richer, moderate-density key distributions but struggling on adversarial distributions.

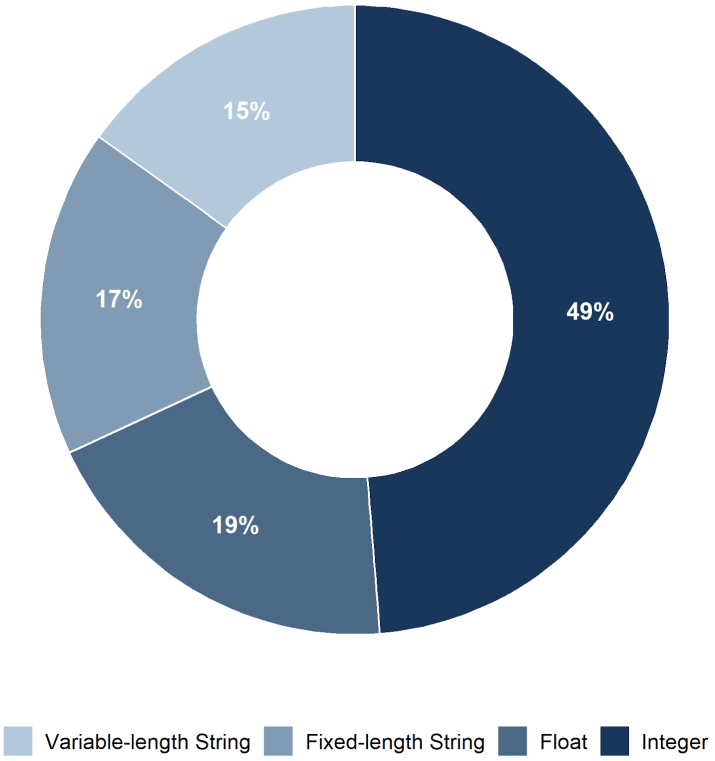


Figure 8. Distribution of Key Types Evaluated Across Learned Index Studies

Table 6. String-key encoding strategies: approaches, performance ranges, and failure conditions.

Strategy	Representative Designs	Throughput / Speedup Range	Where It Works	Where It Breaks
Integer Projection	SLIPP [7], SLIN [8], SwiftKV [9], DobLIX [10]	SLIN [8]: up to 2.47×; SLIPP [7]: 1.3× over ART on random data	Short fixed-length keys; uniform distributions	Long variable keys; high-entropy or adversarial inputs
Prefix-Based Grouping	SIndex [11], RSS [12]	SIndex [11]: 89% throughput gain over XIndex [2] (9:1 read-write); RSS [12]: 7×-70× smaller than ART/HOT	Structured keys with shared prefixes (URLs, paths)	High prefix diversity; tree depth grows unbounded

Feature-Vector Models	SIA [13], SMLP [14]	SIA [13]: 2.6× avg throughput (YCSB); SMLP [14]: outperformed by FST 6.0×-6.5× (negative result)	Rich sub-string features; moderate density	Adversarial distributions; SMLP [14] is the only clean negative result in the corpus

2.5.5 Concurrency and Production Readiness

Fifty-nine studies implement concurrency mechanisms, with six distinct models identified. Reported approaches included epoch-based reclamation for lock-free reads, fine-grained latching at the model-segment level, copy-on-write with version pointers, and single-writer/multiple-reader locks.
Industrial system-integration evidence was reported in only 1 study: Abu-Libdeh et al. integrated a learned index into Google Bigtable's serving stack and reported a 36% reduction in mean point-lookup latency (and 22% for scans) under controlled experimental evaluation on Bigtable's infrastructure, not independent live-production traffic monitoring. This is the strongest system-integration evidence in the corpus, but it should not be read as production-traffic validation at scale. The remaining studies evaluated only in controlled benchmark environments.

Table 7. Concurrency models across the corpus: designs, thread ceilings, and saturation patterns

Concurrency Model	Representative Designs	Thread Ceiling	Saturation Pattern
Fine-Grained Locking	FINEdex [3], LIFT [31], WIPE [69], LITS [26]	16-32 threads (typical)	Monotonic then flat
Coarse-Grained Locking	ALEX [25] (concurrent variant)	8-16 threads	Plateau early
Optimistic Concurrency Control (OCC)	XIndex [2], FINEdex [3], PLIN [16], SALI [51], HIRE [48]	SALI [51]: 64 threads; HIRE [48]: contention past 8 threads	Uneven; dataset-dependent
Lock-Free (CAS / RCU)	LOFT [4], Kanva [5], BLI [49], PFLX [54], DyTIS [71]	LOFT [4]: 80 threads; Kanva/DyTIS: 128 threads	Monotonic to high ceiling
Hardware Transactional Memory	XStore [67]	Platform-dependent	Hardware-limited
GPU Batch Parallelism	G-Learned Index [58]	No thread limit (batch mode)	Synchronisation-free

2.5.6 Seven-Domain Quality Scores

Risk of bias assessment across the seven-domain quality framework revealed concerns. D1 (Baseline Fairness), D2 (Dataset Realism), D3 (Implementation Transparency), D4 (Experimental Execution), D5 (Reporting Completeness), D6 (Performance Measurement), and D7 (Selective Reporting) were evaluated, highlighting variations in how studies approach baseline comparisons, workload types, code availability, failure coverage, reporting of update latency, performance variance, and negative results.

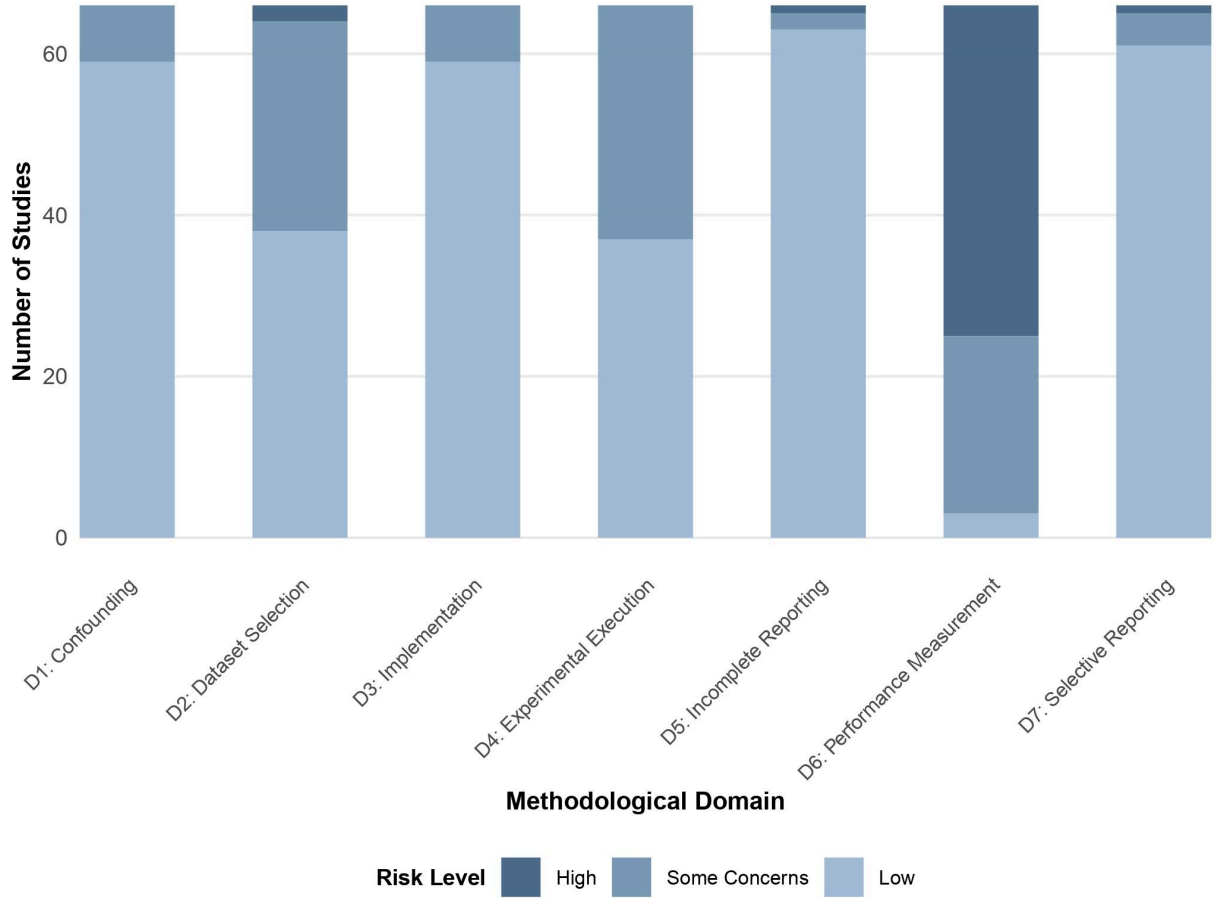


Figure 9. Grouped Percentage Chart Showing the Proportion of Studies with Low Risk, Some Concerns, and High Risk Across Different Bias Domains (Learned Index SLR).

Note: For display, Moderate and Serious were grouped as “Some Concerns,” and Critical as “High Risk.” Domain-level ratings in Table 8 use the full Low / Moderate / Serious / Critical scale defined in Section 1.3.3.

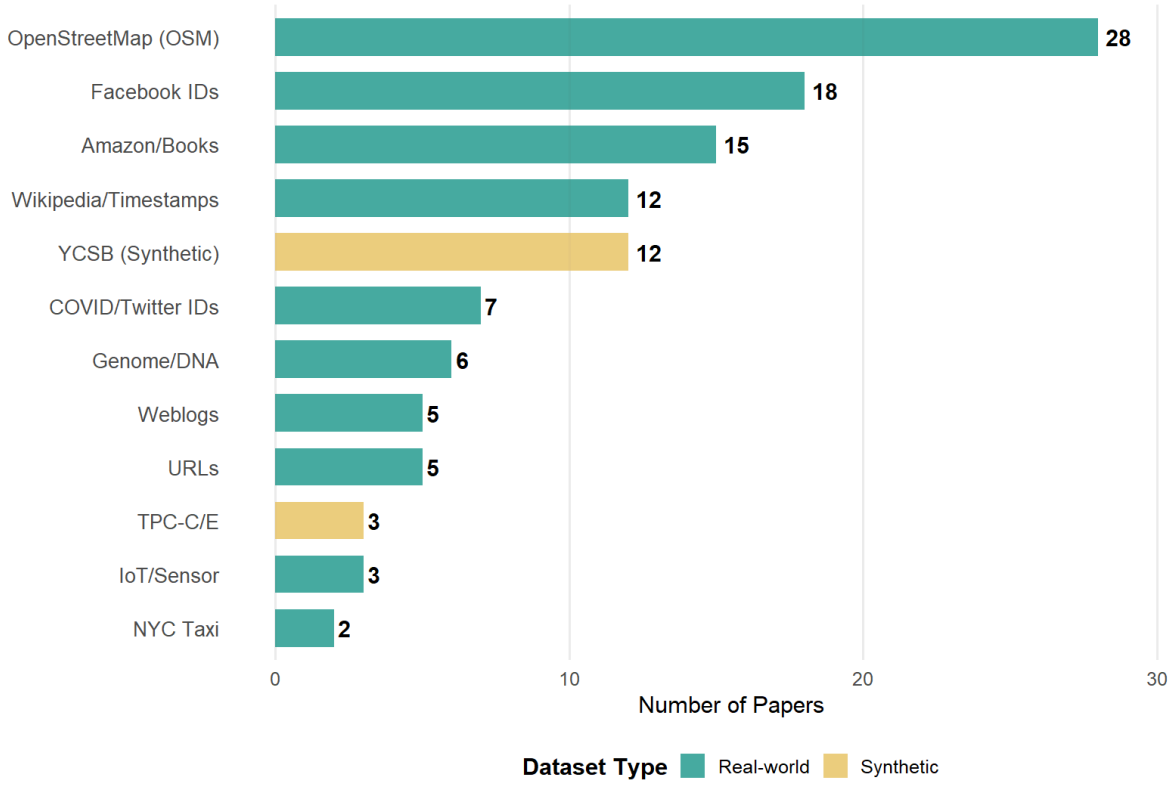


Figure 10. Most Frequently Used Datasets in Learned Index Research.

OpenStreetMap, Facebook IDs, and Amazon Books together account for over 60% of all dataset usages, a concentration directly relevant to the Domain D2 (Dataset Realism) scores below.

Table 8. Risk-of-bias summary across seven methodological quality domains (n=65).

Domain	Focus	Low	Moderate	Serious	Critical
D1: Baseline Fairness	Experimental control & baseline parity	58	7	0	0
D2: Dataset Realism	Distribution diversity & workload coverage	38	25	2	0
D3: Implementation	Code availability & reproducibility	58	7	0	0

D4: Experimental Execution	Failure coverage & concurrency reporting	37	28	0	0
D5: Reporting Completeness	Metric coverage & limitation disclosure	62	2	1	0
D6: Performance Measurement	Variance, warm-up & CI reporting	3	22	35	5
D7: Selective Reporting	Result completeness & reproducibility	60	4	1	0

2.5.7 GRADE Certainty of Evidence

GRADE assessment yielded LOW certainty for all five pre-specified outcomes: point lookup performance (PO1), space and memory efficiency (PO2), range query performance (SO1), update performance (SO2), and concurrency scaling (SO3). The LOW rating across all five outcomes does not reflect a shortage of evidence but its origin: all 65 included studies were proposed and evaluated by the same research team responsible for the system under test, an inherent consequence of inclusion criteria that required novel index proposals. This produces structural positive-result bias, not because the papers are dishonest, but because a field where every performance claim comes from its own authors, with no external replication study inside the corpus, cannot yet be considered settled evidence. Accordingly, every outcome was downgraded for study limitations (self- evaluation) and publication bias, with no upgrade reasons applied to any outcome.

Point-lookup performance (PO1, N=59) showed a consistent benefit, with a median speedup of roughly 2– 4× over B-tree in Regime A and up to 174× in Regime D, but still received LOW certainty because of the downgrades above. Space and memory efficiency (PO2, N=53) also showed a consistent benefit, ranging from 25% to 10,000× savings over B-tree depending on regime, with RUSLI[61] as an exception requiring 3– 4× more memory, yet was likewise rated LOW. (Figure 10. Memory Footprint versus Lookup Latency for Learned and Classic Indexes illustrates this tradeoff: learned indexes push the Pareto frontier toward both smaller memory footprint and lower latency.) Range query performance (SO1, N=34) showed a consistent seek advantage for ranges of 1–100 keys that diminished beyond 1,000 keys, and received LOW certainty for the same reasons. Update performance (SO2, N=46) showed a consistent benefit of 1.06×–32×, with P99 tail latency reduced by up to 10×, but was rated LOW. Concurrency scaling (SO3, N=33) showed monotonic scaling in 32 of the 33 studies addressing it, with gains of 1.5× up to three orders of magnitude over coarse-grained baselines, and was also rated LOW.

Table 9. GRADE certainty summary across five pre-specified outcomes.

Outcome	N	Direction & Magnitude	Downgrade Reasons	Upgrade Reasons	Certainty
PO1: Point Lookup	59	Consistent benefit; median ~2–4× (Regime A), up to 174× (Regime D) exceeding two orders of magnitude in Regime D	Study limitations (self-evaluation); Publication bias	None	LOW
PO2: Space/Memory	53	Consistent benefit; 25%–10,000× vs B-tree (regime-dependent); RUSLI exception (3–4× more memory)	Study limitations; Publication bias	None	LOW
SO1: Range Query	34	Consistent seek advantage for 1–100 key ranges; diminishes >1,000 keys	Study limitations; Publication bias	None	LOW
SO2: Update Performance	46	Consistent benefit; 1.06×–32×; P99 tail latency reduced up to 10×	Study limitations; Publication bias	None	LOW
SO3: Concurrency	33	Monotonic scaling in 32/33 studies; 1.5×–3 orders of magnitude over coarse-grained baselines	Study limitations; Publication bias	None	LOW

2.6 Discussion

2.6.1 Implications For Learned Index Practitioners

The results of current learned index designs are not ready for general-purpose production deployment. The LOW evidence certainty, combined with the absence of crash consistency evaluations, concurrency scalability evidence, and distribution-drift robustness testing, means that adoption in production systems carries substantial uncharacterized risk [74]. Practitioners should: (1) evaluate candidate designs on their specific hardware and workload; (2) require co- located baseline comparisons; (3) assess crash consistency and concurrency behaviour; and (4) plan for distribution-drift monitoring. LSM-tree integration represents the most promising near-term path to production deployment.

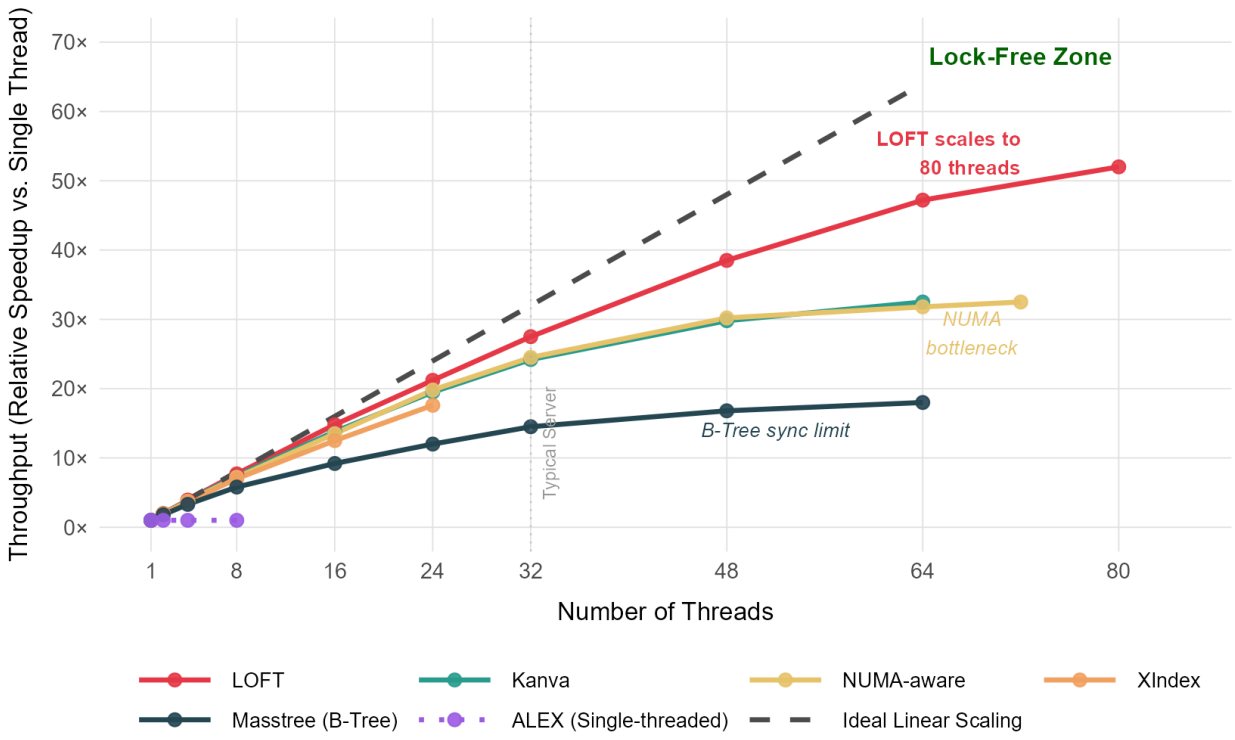


Figure 11: Multi-threaded throughput scaling of concurrent learned indexes. Lock-free designs (LOFT, Kanva) achieve near-linear scaling while traditional B-Trees hit synchronization bottlenecks.

2.6.2 Limitations

Several limitations qualify the findings above. First, the review is scoped to one-dimensional, ordered learned indexes; multi-dimensional, spatial, and learned secondary-index structures were out of scope and

are not represented in any of the conclusions drawn here (see Al-Mamun et al. [75] for a survey of multi-dimensional learned indexes). Second, all 65 included studies were proposed and evaluated by the same team that designed the method, with no independent replication study inside the corpus (Section 1.3.4); this is the principal reason all five GRADE outcomes were rated LOW certainty, and it means the reported speedups should be read as author-reported ceiling performance rather than independently confirmed results. Third, the seven-domain quality scoring (Section 1.3.3) was conducted without a formally reported inter-rater reliability check on an independently rescored subset, so the domain-level risk ratings should be treated as a single-coder assessment rather than a consensus-validated one. Fourth, the search was restricted to English-language publications across seven databases (Section 1.3.2), reference lists were not hand-searched, and grey literature coverage was limited to arXiv, so non-English and unindexed work is not represented. Fifth, the database search strings (Table 1) combine a learned-index core term with both a dynamic/update/lookup modifier and a string-key/persistent-memory/concurrency modifier in a single conjunctive query; a study focused solely on static, numeric, single-key point-lookup performance without also matching one of the string-key/persistent-memory/concurrency terms could therefore have been missed structurally, independent of its relevance. Sixth, the corpus mixes peer-reviewed conference and journal publications with preprint-server sources (Table 3) without stratifying the quality scoring or GRADE certainty assessment by peer-review status; whether preprint inclusion materially affects the reported quality patterns was not separately tested. Finally, studies in the corpus were evaluated on heterogeneous hardware, workloads, and dataset scales that this review did not attempt to normalise, which limits the comparability of raw performance figures reported across studies even where the underlying design comparisons are informative. Finally, as noted in Section 1.3.1, protocol registration occurred after search execution and data collection had already begun rather than in advance of the review, and the review's analytical focus shifted during the work; this retrospective registration and protocol deviation are disclosed explicitly rather than implied to be prospective.

Chapter 3: Scoping Review on Machine Learning for Flood Inundation Mapping

3.1 Introduction

3.1.1 Background

Flooding is among the most frequent and geographically widespread natural hazards, and satellite Earth observation has become central to how flood extent is detected, monitored, and reported in both research and operational emergency- response settings. Within the Earth observation toolkit, the European Space Agency's Copernicus Sentinel-1 and Sentinel-2 missions have emerged as the two most widely used freely available sensors for flood mapping [75]–[78]. Sentinel-1 carries a C-band Synthetic Aperture Radar (SAR) that penetrates cloud cover and operates day or night, making it uniquely suited to capturing flood extent during the very conditions, storms, persistent cloud, low light, under which floods most often occur [75]. Sentinel-2 carries a multispectral optical instrument that offers finer spectral discrimination of water, vegetation, and land cover, but is fundamentally limited by its dependence on clear-sky acquisition [77], [78]. This complementarity has motivated a large and rapidly growing body of research proposing to

fuse S1 and S2 imagery, at the pixel, feature, or decision level, under the assumption that combining the two sensors should outperform either one used alone [76].

Despite the intuitive appeal of this premise, the empirical literature testing it is fragmented. Studies vary widely in the machine-learning and deep-learning architectures used, from classical unsupervised and feature-based classifiers [75], [80] to convolutional encoder–decoder networks [81], decision-level combination schemes [82], and multi-branch feature-fusion architectures [83], and, more recently, transformer- and foundation-model-based approaches [84], [85]. Studies also vary in how "fusion" is operationalized, and, critically, in whether they report a genuine matched-condition comparison against single-sensor baselines or simply assert that fusion is beneficial without isolating the single-sensor performance it is being compared against. A smaller subset of recent studies has additionally reported an unexpected and underexplored pattern: that the benefit of joint training is not symmetric across modalities, with the SAR branch of a fused model sometimes gaining substantially from joint training while the optical branch gains little or even underperforms its standalone counterpart [108]. If this asymmetry is a genuine and generalizable phenomenon rather than an isolated artifact of individual studies, it would have material implications for how fusion architectures should be designed and evaluated going forward, yet no synthesis currently existed, prior to this review, to establish how widespread this pattern is across the broader literature.

A second, related gap concerns validation rigor. Much of the reported "fusion improves performance" evidence is generated and tested within the same flood event or the same benchmark split that the model was trained on. Whether these gains persist when a model is evaluated on a genuinely unseen region, event, or dataset, the condition that matters for real-world deployment, is reported far less consistently. Some recent studies do explicitly test cross-region or cross- dataset transfer [87], [88], including studies that specifically examine how model performance shifts across land-use categories and geographic domains [89, 148], but this remains the exception rather than the norm across the wider literature.

A third gap is operational: Sentinel-2's principal weakness, cloud obstruction, is also the condition under which flood mapping is often most urgently needed. How the literature handles this constraint varies considerably, from lightweight soft-fusion schemes designed around cloud-affected pixels [90], to generative SAR-to-optical translation approaches that reconstruct a synthetic cloud-free view [91], gap-filling frameworks that substitute SAR observations for cloud- obstructed optical scenes [92], and adaptive, cloud-aware sensor-selection frameworks [93]. Yet this handling is rarely synthesized as a distinct axis of comparison across studies, even though it speaks directly to whether reported fusion benefits are achievable in a live flood event rather than only under retrospective, cloud-favorable research conditions.

Several narrative and systematic reviews have addressed SAR-based, optical- based, or multi-sensor flood mapping more broadly, but none has, to this review's knowledge, restricted its scope specifically to S1–S2 fusion and organized the evidence explicitly around matched-condition single-sensor-versus- fusion comparisons, generalization testing, and cloud-cover handling as a joint set of questions. Given the volume and heterogeneity of this literature, and its recent acceleration, with a large share of eligible studies published within the last two to three years, including work explicitly framed around global-scale, near-real-time operational deployment [113], [95], a scoping review following PRISMA-ScR guidance is an appropriate method for mapping the landscape of architectures and fusion strategies, characterizing the

strength and consistency of the evidence that fusion helps, and identifying the methodological and operational gaps that limit translation from research benchmarks to operational flood response systems.

3.1.2 Problem Statement & Scoping Rationale

While individual studies report strong performance metrics, the aggregate evidence base has not been systematically evaluated for methodological quality, generalization capability, or operational readiness. Previous systematic reviews cover adjacent ground, but none combine the three defining features of the present review: Sentinel-1 and Sentinel-2 platform specificity, machine learning or deep learning specificity, and a focus restricted to inundation mapping rather than susceptibility mapping or flood forecasting. Specifically, existing reviews in the literature leave notable gaps:

Bentivoglio et al. (2022) covers deep learning for flood mapping but does so across all sensor systems, lacking platform-specific depth for the Sentinel constellation.

Amitrano et al. (2024) reviews SAR-based flood detection and public datasets but excludes optical data and Sentinel-2 configurations entirely. Destefanis et al. (2025) provides a broad review of Earth Observation services and artificial intelligence techniques, but adopts a narrative rather than systematic methodology.

Manyaka et al. (2026) evaluates remote sensing and machine learning for flood modeling but focuses primarily on applications in developing regions. Silwal et al. (2026) represents the closest overlap by focusing on machine learning and deep learning for flood inundation methods specifically, but fails to evaluate the evidence base through a validation and generalization lens. To address these gaps, this study conducts a scoping review. Unlike a conventional systematic literature review designed to aggregate quantitative effect sizes or perform risk-of-bias assessments, a scoping review is designed to map a heterogeneous evidence base, clarify definitions, and chart the boundaries of research. This review is structured around four research questions. RQ1 maps the landscape of architectures and fusion strategies used to combine S1 and S2 for flood mapping between 2016 and the present, and how this has shifted with the emergence of transformer and foundation-model approaches. RQ2, the spine of this review, asks whether the reported performance gain from fusion, across studies reporting a matched-condition comparison against single- sensor baselines, is consistent, or whether it is architecture-, modality-, and dataset-dependent, including whether the emerging evidence of asymmetric benefit across S1 and S2 branches generalizes across the wider literature. RQ3 examines how these comparisons are validated, same-event versus cross-region or cross- dataset testing, and how much of the reported fusion benefit survives generalization testing outside training conditions. RQ4 asks how fusion studies handle Sentinel-2 cloud unavailability at the time of a flood event, and what this implies about real-world deployability versus retrospective research performance.

3.1.3 Aim and Scope Boundaries

By synthesizing evidence from 73 eligible studies published between 2019 and 2026, this review provides a focused, PRISMA-ScR-guided account of whether, and under what conditions, S1–S2 fusion measurably improves flood inundation mapping, and identifies the specific evidentiary and

methodological gaps that stand between current research performance and operational deployment. To maintain high rigor, the review operates under three explicit, non-negotiable scope boundaries:

1. Satellite Platform Limitation: The evaluation is strictly limited to studies utilizing Sentinel-1 and/or Sentinel-2. Studies relying solely on other satellite missions (such as Landsat, MODIS, or commercial high-resolution constellations) are excluded.
2. Multi-Source Fusion Limitation: Sensor fusion is defined strictly as the integration of Sentinel-1 SAR data with Sentinel-2 optical data. Studies fusing satellite imagery with non-satellite sources (such as social media, crowd- sourced data, or physical hydraulic models) are excluded from the fusion- specific synthesis.
3. Task Limitation: The scope is restricted to flood inundation detection, segmentation, mapping, or monitoring. Studies focusing on flood susceptibility mapping, long-term forecasting, or post-flood damage/infrastructure assessment are excluded.

3.2 Scoping Review Methodology

3.2.1 PRISMA-ScR Framework and Protocol

This scoping review was conducted following the Joanna Briggs Institute (JBI) methodology for scoping reviews and is reported in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR), utilizing its 22-item checklist to ensure comprehensive and transparent reporting. In accordance with PRISMA-ScR guidelines and the nature of scoping reviews as evidence-mapping tools, no formal risk-of-bias assessment or GRADE certainty rating was conducted; critical appraisal remains optional for scoping reviews, as the primary objective is to map the volume, nature, and characteristics of the primary research rather than synthesizing highly standardized clinical or experimental interventions. Protocol was registered in advance of conducting this review in OSF (Double Blind Registration). The charts and tables used in this review are available in GitHub Repository (https://github.com/mdasifbinkhaled/Flood-SLR).

3.2.2 PCC Framework

To guide study selection and ensure alignment with the review's core objectives, the review was structured using the Population, Concept, and Context (PCC) framework:

1. Population: Flood events or flood-prone areas, of any flood type (fluvial, pluvial, coastal, or flash), captured by Sentinel-1 and/or Sentinel-2 imagery, in any geography.
2. Concept: Studies applying a machine learning or deep learning method to flood inundation detection, segmentation, mapping, or monitoring, where either (a) Sentinel-1 and Sentinel-2 imagery were jointly used as direct fused input to the model's inference pipeline at the pixel, feature, or decision level, or (b) the study reported a matched-condition empirical comparison between a single-sensor (Sentinel-1-only or Sentinel-2-only) model and a Sentinel-1-plus-Sentinel-2 fusion model using the same dataset, event, or architecture.
3. Context: No geographic restrictions (global coverage). All flood typologies are eligible, including riverine, pluvial, coastal, and flash flooding. The temporal coverage spans from 1 January 2016 to

1 September 2026, the final search date. The 2016 start does not coincide with the missions' launches (Sentinel-1A: April 2014; Sentinel-2A: June 2015); it was chosen as a practical cutoff after both constellations had entered routine operational data delivery, rather than to mark either launch date. Only peer-reviewed, English-language publications, or non-English publications with an available English translation, are included.

3.2.3 Eligibility Criteria and Information Sources

Studies were excluded if Sentinel imagery was used only for ancillary or inventory purposes with no fusion input and no matched comparison; if the study used only a single sensor with no fusion comparison of any kind; if the method was purely threshold- or rule-based with no machine learning or deep learning component (unless used as a comparator to an eligible method); if no empirical evaluation was reported; or if the study addressed flood susceptibility mapping, hydrological or discharge forecasting, damage assessment, or permanent water extraction rather than inundation detection and mapping. Publication type was restricted to peer-reviewed journal articles, conference papers and preprints; theses, and magazine articles were excluded, and where a conference paper was later extended into a full journal version, only the more complete journal version was retained to avoid double-counting the same experiment. Inclusion required a quantitative performance metric (for example, intersection over union, mean intersection over union, F1 score, overall accuracy, Cohen's kappa, precision, recall, or critical success index) or documented operational performance; a small number of studies reporting only a qualitative or visual single-sensor-versus-fusion comparison, with no numeric metric, were included but flagged as low-confidence evidence rather than weighted equally with quantitatively reported comparisons. Reviews, surveys, and meta-analyses meeting the topical scope were not coded as primary evidence but were tracked separately for citation chasing and for positioning in the Discussion. Searches were conducted across one bibliographic database (Scopus), one academic search engine (Google Scholar), and one discovery/citation tool (Semantic Scholar), supplemented by publisher-platform searches across SpringerLink, IEEE Xplore, ScienceDirect, Taylor & Francis, Wiley, and MDPI. The search string applied was: ("Sentinel-1" OR "SAR") AND ("Sentinel-2" OR "optical") AND (fusion OR "multi- sensor" OR "multi-modal") AND (flood* OR inundation) AND ("machine learning" OR "deep learning" OR neural network*)

Filters restricted results to publication years 2016–2026, document type to journal articles and conference papers, and language to English.

3.2.4 Screening, Data Charting, and Critical Appraisal

Records retrieved from all databases were combined and deduplicated. Title/abstract screening and full-text eligibility were shared across four reviewers. Each record was assessed by at least one reviewer; uncertain or borderline records were independently assessed by a second reviewer and resolved by discussion. Charting (data extraction) was performed by one reviewer per record and spot-checked by a second reviewer on a subset of studies. No formal inter-rater reliability statistic (e.g., Cohen’s κ) was calculated; disagreements were resolved by consensus. Ambiguous-case decision rules were applied consistently during screening: a model trained on fused Sentinel-1 and Sentinel-2 data but evaluated on a single modality at inference was included; rule-based indices combined post hoc with no machine learning component were excluded unless used as a comparator baseline; and foundation models

pretrained on general multimodal data and fine-tuned on Sentinel-1 and Sentinel-2 flood data were included. The full PRISMA-ScR screening flow is reported in Section 2.3.1.

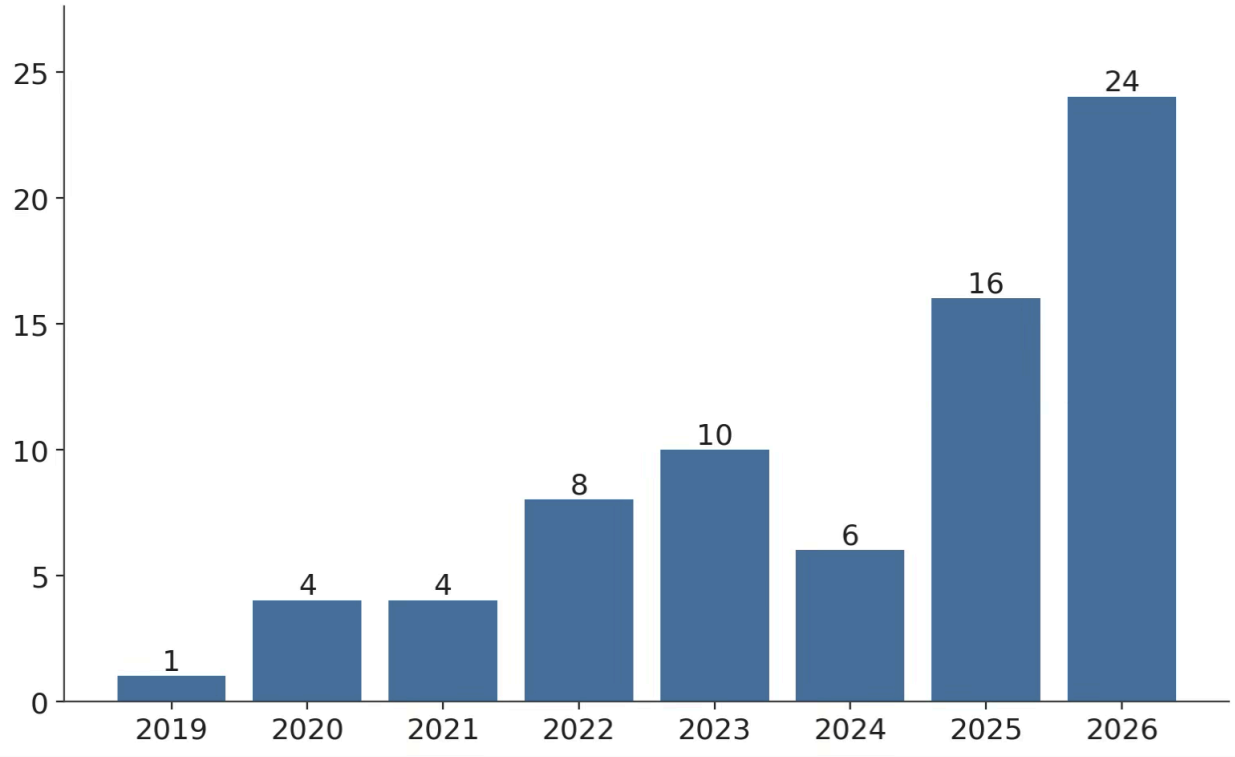


Figure 12: Growth of Sentinel-1–Sentinel-2 fusion flood-mapping literature, 2016–2026 (n = 73 included studies).

Publication year was coded as the year of the final archival version available at the 1 September 2026 cutoff (journal issue year or formal proceedings year). Online-first dates were not used when a later issue/proceedings year existed. Preprints used the arXiv first-posted year unless a peer-reviewed version with a different year was already public at cutoff. Figure 13 is generated from this single rule. A structured data-charting table was developed and piloted on a sample of included studies before full use, and charting was performed independently by four reviewers, with a process for reconciling discrepancies (for example, cross-checking a subset of records or consulting a second reviewer for uncertain extractions). Where information required for a given field was not reported in a source, this was recorded explicitly as "not reported" or "not applicable" rather than left blank, to distinguish an absence of data from an oversight in charting. Data were charted across five groups of variables corresponding to the review's bibliographic and screening information and its four research questions: (1) bibliographic and screening data (serial number, title, authors, year, document type, venue, eligibility category, evidence type, confidence flag, and screening notes); (2) study context (country/region, flood type, event narrative, and study-area or dataset scale); (3) RQ1 items (Sentinel-1/Sentinel-2 product and acquisition details, fusion level, fusion mechanism, auxiliary data, model architecture, training paradigm, dataset(s), and number of events/samples); (4) RQ2 items (whether a matched comparison was reported, metric types and values for each configuration, per-modality branch performance under joint training, authors' stated direction and rationale, and facilitators/barriers to fusion benefit); (5) RQ3 items (validation design, split

strategy, whether cross-region/cross-dataset generalization was tested, and generalization results); and (6) RQ4 items (cloud-cover handling method, discussion of cloud cover as a real-world constraint, temporal offset reporting, and claims or evidence of real-time/operational deployment), alongside synthesis fields capturing authors' stated key findings, limitations, and future work in their own terms. No variables were imputed; where a study reported multiple values for the same metric, all reported values were charted, and the narrative comparison field records how these were interpreted collectively. Consistent with JBI and PRISMA-ScR guidance that formal quality or risk-of-bias appraisal is not a required component of a scoping review, no formal critical- appraisal instrument was applied to individual sources. Instead, a confidence flag was assigned to each source at the point of data charting: standard- confidence evidence, defined as a quantitatively reported matched single-sensor- versus-fusion comparison or a directly reported fusion-input result, was distinguished from low-confidence evidence, defined as a qualitative or visual- only comparison with no numeric metric. This flag was used during synthesis to ensure that low-confidence evidence was interpreted alongside, rather than weighted equally with, quantitatively supported findings, particularly for RQ2. Charted data were synthesized narratively and descriptively rather than through meta-analysis, consistent with the heterogeneity of architectures, datasets, flood events, and metrics expected across the included literature.

3.3 Results

3.3.1 Selection and Characteristics of Included Studies

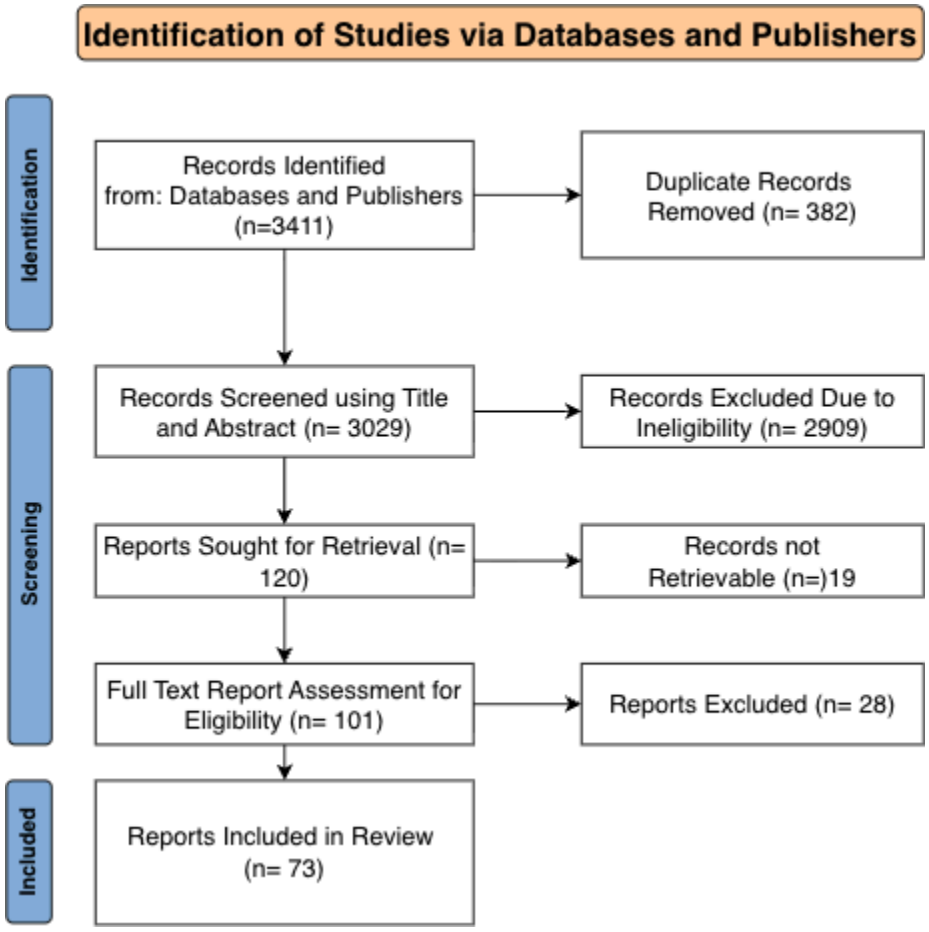


Figure 13. PRISMA Flow Diagram

Database searching identified 3,411 records. After removal of 382 duplicate records, 3,029 titles and abstracts were screened, of which 2,909 were excluded as clearly ineligible. Full texts of 120 records were sought for retrieval, of which 19 were not retrievable, leaving 101 reports whose full text was assessed for eligibility. Of the reports assessed in full text, 28 were excluded. Exclusion was driven by the eligibility criteria in Section 2.2.3 (single-sensor only with no fusion comparison; no empirical evaluation; susceptibility/damage/forecast focus rather than inundation mapping; publication type or date outside scope). A per-reason count ledger was not retained after screening; the reasons above are listed in order of how often they arose during adjudication, but exact frequencies are not reported. After full-text assessment, a total of 73 reports were included in the review. The 73 included sources span publication years 2019 through 2026, with a pronounced concentration in the most recent period: 1 study was published in 2019, 4 in 2020, 4 in 2021, 8 in 2022, 10 in 2023, 6 in 2024, 16 in 2025, and 24 in 2026

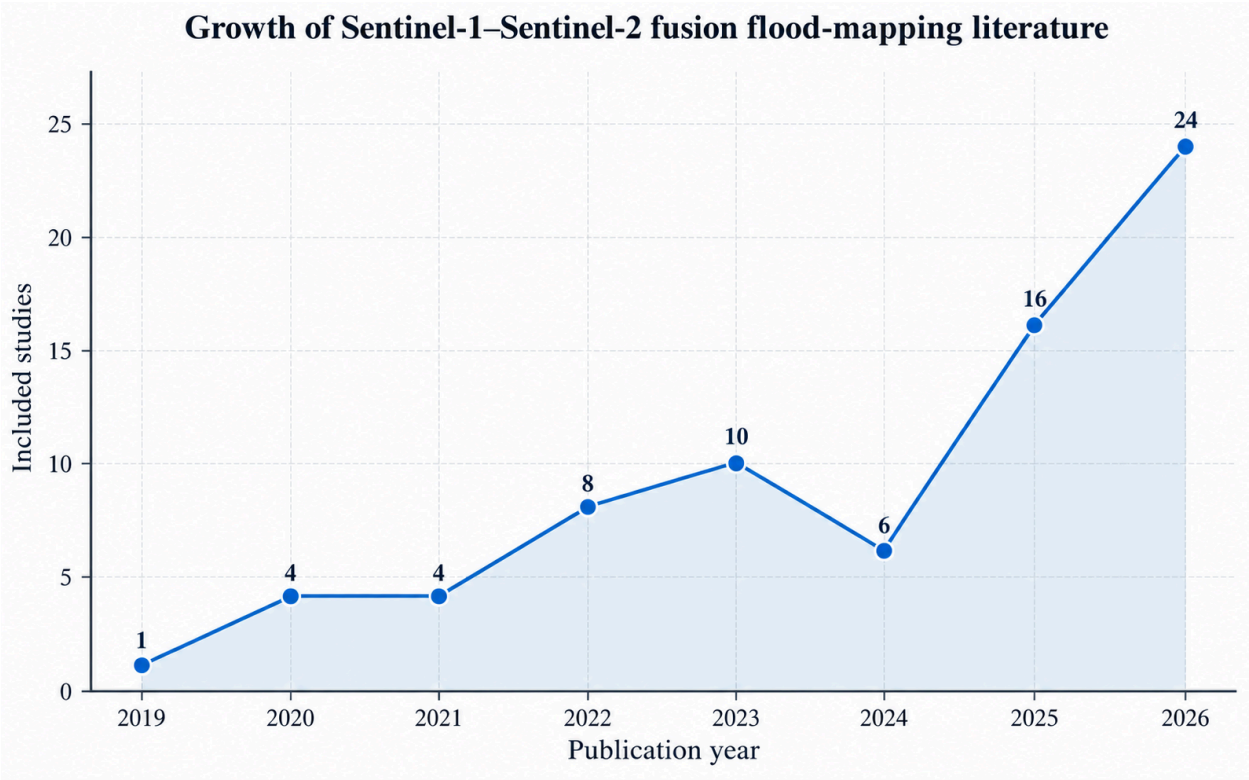


Figure 14. Number of Publications by Year

Geographically, included studies drew on flood case studies spanning Africa, Asia, Europe, North and South America, and Oceania, with individual studies typically covering one to several named events (for example, Rockhampton, Texas, and Kerala [79]; Ordu, Turkey [100]; the Kakhovka dam disaster [111]; Porto Alegre and Maputo [113]; the Baige landslide dam on the Jinsha River [84]; and the Derna dam collapse [139]) alongside a growing number of large, multi- continent benchmark datasets (for example SEN12-FLOOD [125], Sen1Floods11 [76, 86], WorldFloods [104], and GEOID-Flood [119]). Flood types included fluvial, pluvial, coastal, and flash flooding, with the majority of studies addressing riverine or mixed fluvial–pluvial events and a smaller subset focused specifically on flash floods or outburst floods [84].

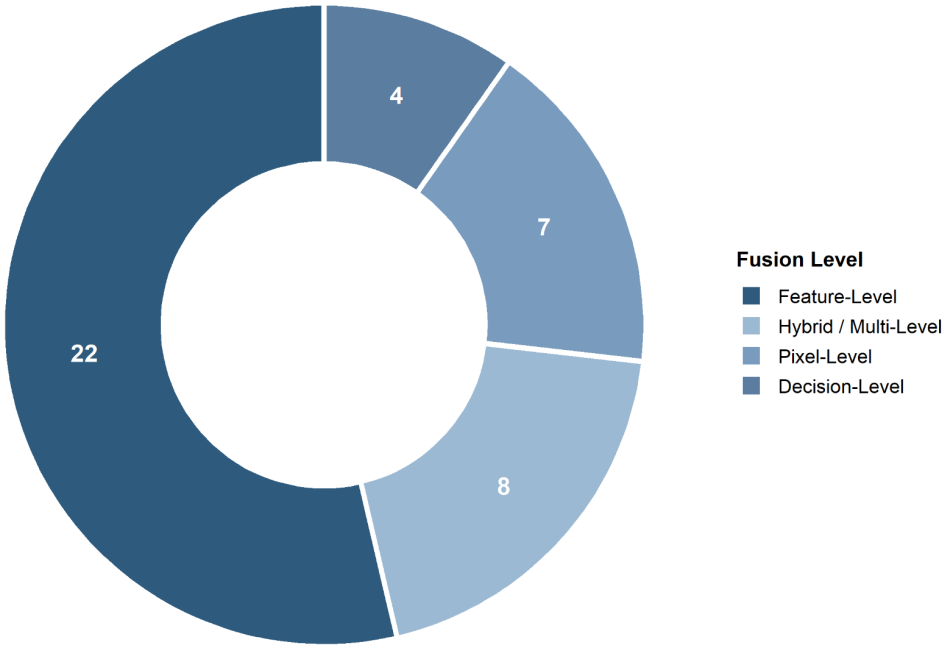


Figure 15. Fusion-level strategy- pixel-, feature-, decision-, or hybrid-level

Under the review's eligibility categories, most sources satisfied both category (a), direct use of Sentinel-1/Sentinel-2 fusion as model input, and category (b), a matched single-sensor-versus-fusion comparison, simultaneously. A smaller subset satisfied only category (a), typically multimodal pretraining or representation-learning studies that fuse the two sensors during training but do not always report an isolated single-sensor ablation [91, 98, 99], and a further subset satisfied only category (b) without treating fusion as the primary architectural contribution [80, 103, 105, 106].

Consistent with the Methods, no formal risk-of-bias instrument was applied; sources were instead flagged descriptively at charting. The great majority of sources (69 of 73) provided quantitative performance metrics, either alone or alongside qualitative or visual comparisons, and were treated as standard-confidence evidence. A small number of sources (4 of 73) reported only a qualitative or visual comparison between single-sensor and fusion configurations, with no numeric metric, and were flagged as low-confidence evidence per the ambiguous-case rule; these sources are retained in the RQ1 landscape synthesis but are explicitly down-weighted in the RQ2 synthesis of whether fusion improves performance. Full study-level data for each of the 73 included sources, including bibliographic details, fusion level, architecture, dataset, matched comparison values, validation design, and cloud-cover handling, are available in the accompanying data-extraction table.

3.3.2 Landscape of Architectures and Fusion Strategies (RQ1)

Architectural choices for combining Sentinel-1 and Sentinel-2 shifted markedly over the review period. Early studies (2019 to 2021) relied on classical machine learning and shallow fusion pipelines, for example a Kohonen self-organizing map applied to hand-engineered texture and water-index features [75], unsupervised clustering approaches [79], and Random Forest or Support Vector Machine classifiers operating on stacked optical, SAR, and auxiliary features [100, 101, 106]. From roughly 2021 onward, U-Net-style convolutional encoder–decoder architectures became the dominant design for pixel-wise segmentation, appearing in various forms across the literature (for example [87, 88, 102, 108, 138, 140, 141]), often as both a baseline and as the backbone for more elaborate fusion modules.

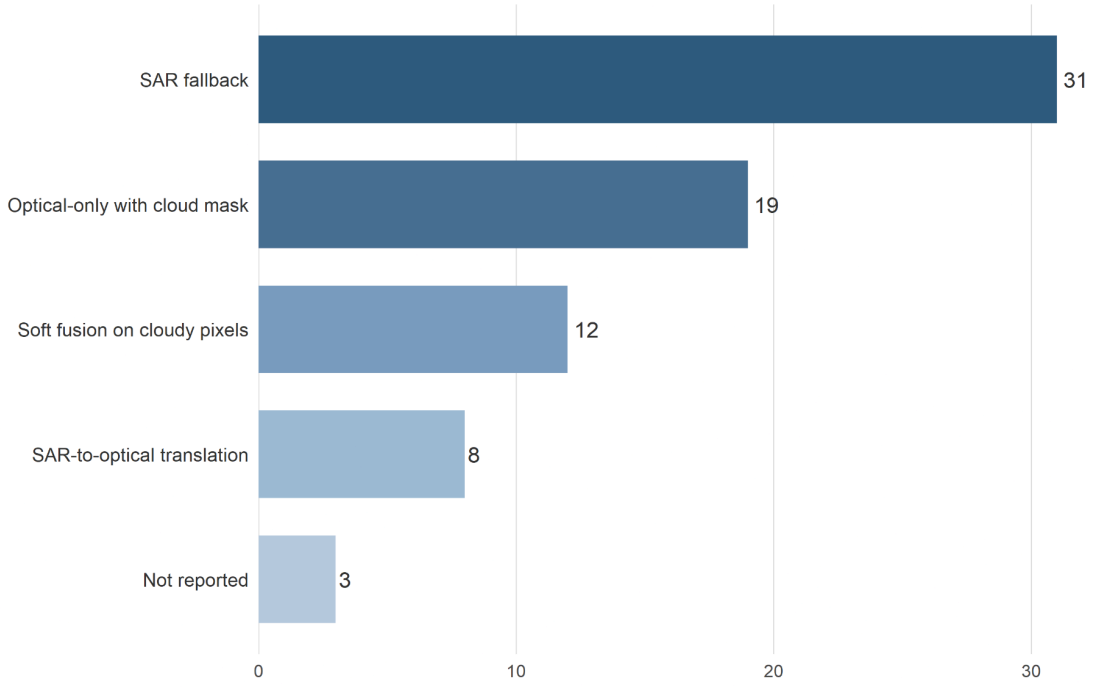


Figure 16: Cloud Handling Methods

Fig. 17 compares the cloud-cover handling strategies used across the 69 studies (of 73 total) that addressed Sentinel-2's cloud limitation, with some studies using more than one approach so the totals don't sum to 69. Cloud masking/filtering and reliance on cloud-penetrating Sentinel-1 SAR as a fallback are tied as the most common strategies, each appearing in 27 studies, followed by temporal compositing/gap-filling and cloud-free scene selection, each used in 14 studies [115]. Far fewer studies adopt more technically advanced solutions: SAR-to- optical synthesis/translation via generative models appears in only 8 studies, and learned dynamic fusion weighting, where attention or gating mechanisms adaptively reweight sensors based on cloud presence, appears in just 3. Overall, the distribution shows that the field still leans heavily on simple, well- established preprocessing techniques rather than adaptive, learning-based fusion strategies.

Feature-level fusion, in which separate Sentinel-1 and Sentinel-2 encoder branches are merged by concatenation, attention, or gating prior to a shared decoder, was the single most common fusion strategy across the corpus, appearing in some form in roughly half of all included studies, while pixel-level (early, input-stacking) fusion and decision-level (late, output-combining or ensembling) fusion were each used in a substantial minority of studies, and a number of more recent architectures combined more than one fusion level within a single pipeline (for example simultaneous feature- and decision-level fusion in [83, 106, 116, 124]). Table 10 summarises the fusion level used by each of the four dominant architecture families across the corpus (N = 73).

From around 2023, and accelerating sharply through 2025 and 2026, the literature shows a clear shift toward transformer, state-space (Mamba), diffusion, and foundation-model architectures. Transformer-based fusion appeared in frequency- domain and cross-modal attention forms [85, 126, 127]; state-space Mamba backbones were adapted for cross-modal flood change detection and cloud-aware fusion [117, 120, 121, 135]; diffusion-based approaches were used both for SAR- to-optical translation and for joint diffusion-convolution fusion [111, 130, 131]; and large pretrained geospatial foundation models, including Prithvi-based encoders, Galileo, and other multimodal masked autoencoders, were increasingly fine-tuned or adapted for flood segmentation rather than trained from scratch [84, 87, 94, 98, 99, 119, 146]. Several 2025 and 2026 studies explicitly built on or benchmarked against these foundation models, for example comparing a fusion- adapted Prithvi variant against the original pretrained encoder [87] and evaluating five geospatial foundation-model backbones under early fusion on a new global benchmark [119].

Auxiliary data beyond Sentinel-1 and Sentinel-2 were incorporated in a substantial minority of studies, most commonly digital elevation models for terrain correction or slope masking [85, 112, 139], precipitation and hydrological indices [111, 119], and very-high-resolution commercial or UAV imagery used for independent validation rather than as model input [139, 147]. Dataset scale varied widely, from small, single- or triple-event case studies with no formal benchmark name [79, 100, 111] to large standardized benchmarks reused across many studies, most prominently Sen1Floods11 [76, 86, 87, 88, 98, 118, 128, 132, 137] and SEN12-FLOOD [125, 133, 142], alongside a growing number of purpose-built large-scale fusion benchmarks introduced in the most recent period, including WIPI [109], S1S2Water [107], GEOID-Flood [119], and IT-SS-18K [139].

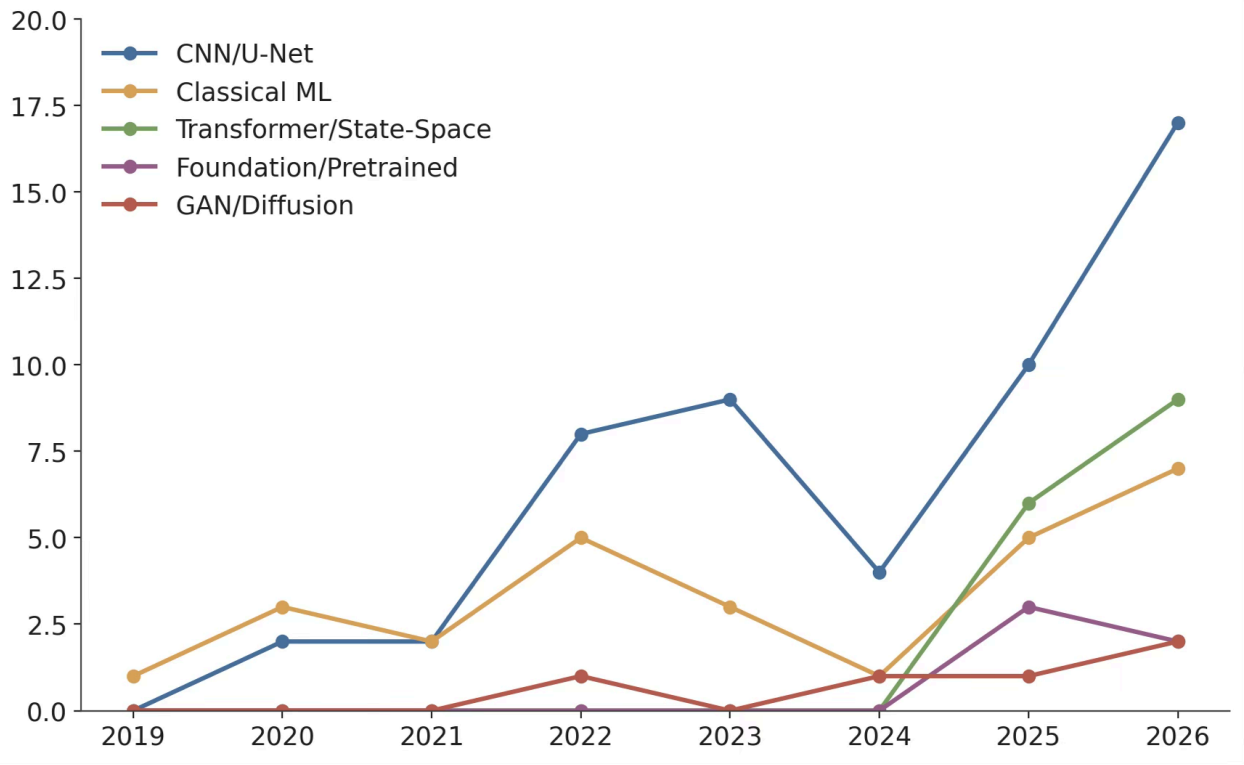


Figure 17. Temporal shift in architecture families, showing transformer and foundation-model approaches emerging only from 2025 onward (RQ1)

3.3.3 Does Fusion Improve Performance? (RQ2)

Of the 73 included sources, 59 reported a matched, same-dataset or same- architecture comparison between at least one single-sensor configuration (Sentinel-1-only or Sentinel-2-only) and a fused Sentinel-1-plus-Sentinel-2 configuration; a further 11 evaluated only fused configurations without an isolated single-sensor baseline, and 3 could not be classified with confidence from the reported text. Among the 59 matched comparisons, the overwhelming majority, 65 of the 73 sources overall when qualitative and low-confidence comparisons are also counted, reported that fusion improved performance relative to the best single-sensor baseline, and only a small number reported a genuinely mixed, asymmetric, or neutral result.

Table 10. Distribution of Reviewed Studies by Model Architecture Family

Architecture Family	Count	Models used
CNN / Encoder-Decoder	29	U-Net, UNet++, DeepLabV3+, BASNet, CMCDNet, MFU-Net
Classical ML / Statistical	18	Random Forest, SVM, Gradient Boosting, k-NN, OBIA rule-based, Markov Random

		Field
Hybrid / Multi-Architecture	12	CNN+SVM, RF+CNN, Transformer+CNN, ResNet+GRU, Physics-Guided UNet-FNO
Transformer / Foundation Model	7	Vision Transformer (Galileo), TerraMind, SegFormer/Swin-T, TLE-FEDformer, SPEAR
Mamba / State-Space	3	SD-Mamba, CISRMamba, M²F-Net (MambaVision)
GAN / Generative	2	cGAN (Pix2Pix) + U-Net, GAN-MP
Diffusion-Based	2	DSE (VQ-GAN + Brownian Bridge Diffusion), JSDFloodNet

To move beyond this directional count and quantify the size of the reported effect, the charted Sentinel-1-only, Sentinel-2-only, and fusion performance values were extracted directly for every source that reported a numeric value for the same primary metric (IoU/mIoU, Overall Accuracy, or F1) across all three configurations. This yielded 45 of the 73 sources with a directly comparable, single-metric, matched-condition data point (31 reporting IoU/mIoU, 7 reporting Overall Accuracy, and 7 reporting F1), summarized in Table 4 below. Across these 45 studies, the fused model outperformed the stronger of the two single-sensor baselines in 37 cases (82.2%), matched it exactly in 2 cases, and underperformed it in 6 cases (13.3%). The mean fusion gain over the stronger single-sensor baseline was +6.2 percentage points and the median gain was +4.2 percentage points (standard deviation 7.8 points; range -3.0 to +29.9 points), confirming numerically what the qualitative synthesis above suggested: a real but highly right-skewed benefit, in which a minority of studies drive a disproportionate share of the apparent average gain. Binning these 45 gains by magnitude, 8 studies showed a decline or effectively no change (≤0 points), 8 showed a gain of 0 to 2 points, 8 showed a gain of 2 to 5 points, 11 showed a gain of 5 to 10 points, 6 showed a gain of 10 to 20 points, and 4 showed a gain exceeding 20 points [81, 102, 105, 133].

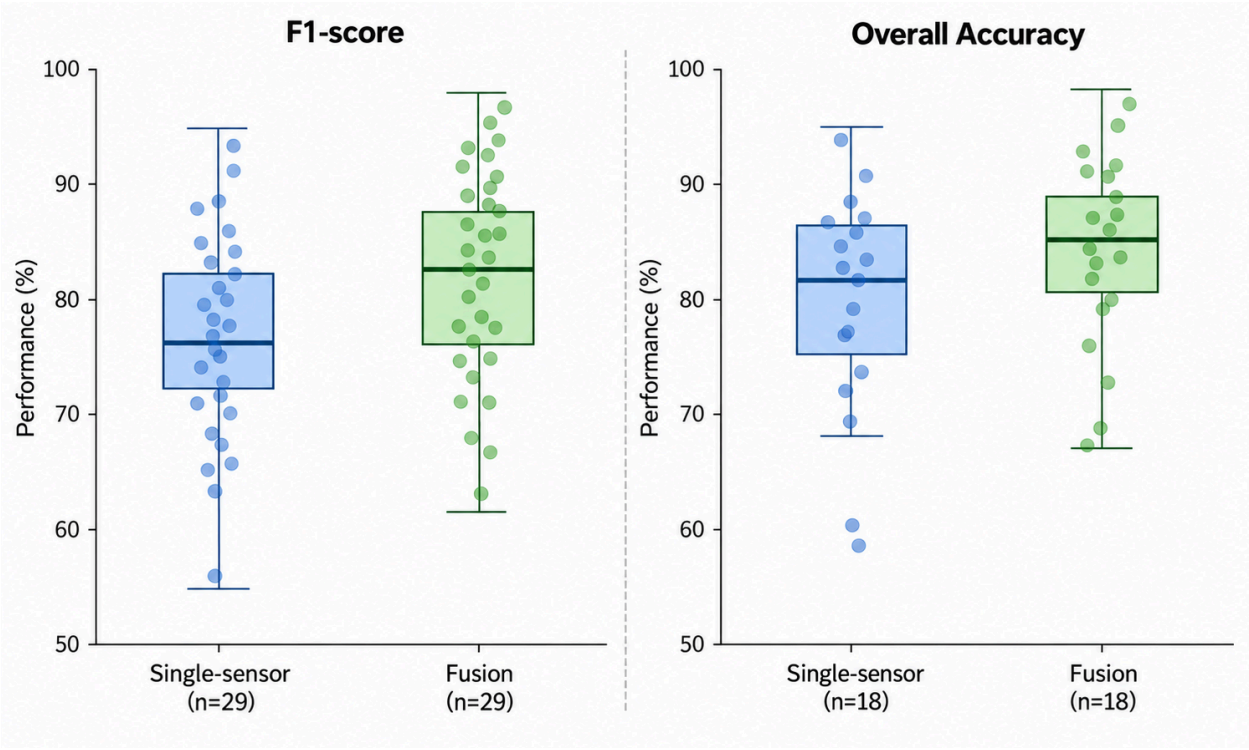


Figure 18. Comparison of F1-Score and Overall Accuracy Across Single-Sensor and Sentinel-1–Sentinel-2 Fusion Approaches Grouped by Evaluation Metric

The size of the reported fusion gain, however, varied enormously by architecture, dataset, and metric, which speaks directly to the review's central question. At one extreme, several studies reported fusion gains of ten to twenty or more accuracy or IoU points over the stronger single-sensor baseline: an early channel-stacked Sentinel-1-plus-Sentinel-2 U-Net raised water-class overall accuracy from a Sentinel-1-only 57.75% (or Sentinel-2-only 92.73%) to 95.14% for the fused Siamese model [137]; a cloud-guided Mamba fusion model improved F1 from roughly 0.83 for the better single-sensor branch to 0.93 for the fused model [135]; a cross-modal Mamba architecture raised mean IoU from a single-sensor best of about 61.6% to 91.5% for the fused configuration [133]; and a frequency-domain transformer fusion model reported an IoU improvement from roughly 90% for the best single-sensor baseline to 97.4% for the proposed multi- sensor model [124].

A number of mid-range studies illustrate the middle of this distribution and show that fusion gains were not confined to any single architecture family. A classical Random Forest classifier fusing SAR, multispectral, and terrain features raised overall accuracy from a Sentinel-1-only 77.8% (IoU 0.45) and a Sentinel-2-only 84.1% (IoU 0.35) to a fused 88.5% (IoU 0.62) in a Himalayan mountainous basin [144]. A gradient boosting model fusing Sentinel-1, Sentinel-2, and DEM features improved overall accuracy from 85.2% (SAR-only) and 89.6% (optical-only) to 93.75% for the two-sensor fusion and 96.25% once DEM was added, with Kappa rising from 0.70 (SAR-only) to 0.93 (full fusion) [126]. A multi-branch U-Net fusing SAR and water-index features raised F1 from 77.7% (SAR-only) and 89.7% (best single optical configuration) to 91.9%, with mean IoU improving from 63.2% to 85.0% [96]. A hybrid CNN-SVM time-series fusion model raised overall accuracy from 82.5% (SAR-only) and 85.0% (optical-only) to 93.8%, with Kappa improving from 0.65-0.70 to 0.865 [68]. Feature-space fusion of

SAR polarization and optical water indexes using a Support Vector Machine raised overall accuracy from 84.0% (SAR-only) to 93.0% (fused SAR plus one optical water index) and up to 98.0% when a stronger optical index was used, with Kappa rising from 0.60-0.63 for single-sensor configurations to 0.97 for the best fused configuration [129]. At the other extreme, a comparable number of studies reported gains of only one to three percentage points, or in a few cases essentially no improvement or a small decrease relative to the stronger single sensor branch. A U-Net++ style architecture found that adding Sentinel-1 to a strong Sentinel-2-only baseline changed water IoU only marginally, from about 90 to 91% depending on configuration [141]; a masked autoencoder pretraining comparison found the fused variant outperformed a Sentinel-2-only baseline by well under one percentage point on downstream mean average precision [99]; and one U-Net based multimodal water segmentation study reported that a Sentinel-2- only branch already reached 94.86% IoU, with the fused model performing essentially the same, at 94.72%, that is, marginally lower [86]. A small number of studies reported an outright negative result for at least one configuration, most notably an early fusion Sentinel-1 plus Sentinel-2 comparison in which the fused IoU (65.0%) fell below the stronger single sensor, Sentinel-2-only, baseline (68.0%) [93, 119], and a large benchmark study of foundation model backbones in which fused overall accuracy (92.3%) fell 2.7 points below the Sentinel-2-only NDVI threshold baseline (95.0%) despite outperforming the Sentinel-1-only baseline (77.7%) by nearly 15 points [83].

Table 11. Distribution of Sentinel-1/Sentinel-2 Fusion Studies by Fusion Level and Single-Sensor Baseline Reporting

Fusion Type	S1-only tested	S2-only tested	S1+S2 (fusion) tested
Pixel-Level (Early) Fusion	11	10	12
Feature-Level (Intermediate) Fusion	33	30	39
Decision-Level (Late) Fusion	5	4	6
Hybrid / Multi-Level Fusion	14	14	14
Other / Not Specified	2	2	2

This heterogeneity was strongly patterned by which single-sensor branch was already strong for a given flood type and land cover. Where Sentinel-2 was severely degraded, most often by cloud cover over the flood event, or where SAR alone struggled with vegetated or urban flooding, fusion gains tended to be large [85, 120, 121, 145]; where both single-sensor branches were already close to saturation on a relatively easy benchmark, for example open-water detection on cloud-free imagery, fusion gains tended to be small or negligible [86, 99, 141]. Several authors made this pattern explicit, for instance noting that fusion primarily resolves cases where SAR misclassifies vegetation-obscured or soaked ground and optical imagery fails over open water under cloud cover or specific illumination conditions [125], or that a

proposed decision-level or hybrid fusion mechanism narrowed the gap specifically in scenes where the single-sensor branches disagreed [116, 122, 124].

Table 12. Benchmark dataset usage across reviewed studies

Benchmark Dataset	Studies Using It
Sen1Floods11	29
SEN12-FLOOD	7
WorldFloods	6
CAU-Flood	4
SEN12MS	4
S1S2-Water	3
Kuro Siwo	2
UNOSAT	2
OMBRIA	1
Custom / no named public benchmark	20
Any named public benchmark (≥1)	53

The review's protocol specifically asked whether branch-level asymmetry (one modality improving under joint training while the other does not) generalizes across the wider literature. Twenty-two of the 73 sources reported per-modality branch performance separately under joint or fused training, allowing this question to be examined directly. In most of these studies, both the Sentinel-1 and the Sentinel-2 branch were reported to improve under joint training, with authors attributing this to complementary error correction, for example SAR resolving optical cloud or shadow ambiguity while optical resolves SAR speckle or vegetation ambiguity [85, 87, 88, 118, 119, 121, 124, 140, 141, 142, 144]. However, a genuinely asymmetric pattern was also documented in a meaningful minority of these studies. One U-Net comparison found that a Sentinel-1 branch improved substantially under fusion while the paired Sentinel-2 branch showed only a slight decrease or effectively no change [86]. A cross-modal contrastive pretraining study similarly reported that the Sentinel-1 encoder improved consistently under joint pretraining, whereas the Sentinel-2 encoder improved only in data-scarce regimes and showed comparable, not improved, performance once more labeled data became available [98]. A separate multi-branch water- segmentation study reported the reverse pattern in one configuration, where naive early channel-stacking caused the fused network to rely predominantly on the optical channels, leaving both single-sensor contributions effectively unchanged relative to a stronger fusion mechanism evaluated in the same paper [89]. Together, these findings indicate that asymmetric branch benefit is a real and recurring phenomenon in the literature, but it is not universal, and its direction, whether Sentinel-1 or

Sentinel-2 benefits more, or whether one branch is entirely absorbed by the other, itself depends on architecture (particularly how the two branches are weighted or gated) and on flood or land-cover context, rather than following a single consistent rule.

Reported facilitators of fusion benefit across the corpus included attention or gating mechanisms that adaptively reweight sensor contributions under cloud cover or terrain complexity [85, 120, 121], auxiliary DEM or physics-informed constraints that reduce ambiguity in flooded terrain [122], and temporal or multi-date compositing that allows the two sensors to compensate for each other's revisit gaps over time [92, 125]. Commonly reported barriers included naive concatenation strategies that allow one modality to dominate the learned representation [89], a persistent scarcity of large, precisely co-registered, densely labeled Sentinel-1-plus-Sentinel-2 training sets relative to single- sensor archives [125], and reduced or negative fusion benefit on benchmarks where the easier single-sensor modality already performs close to ceiling [86, 99].

Table 13. Distribution of Reviewed Studies (n = 73) by Reporting of Sentinel-2 Cloud-Cover Handling Method and Claimed Operational/Real-Time Deployment

Category	Count	% of corpus
Both mentioned (cloud-cover handling addressed AND operational deployment claimed)	43	58.9%
Only one mentioned (cloud-cover handling addressed, but no deployment claim or vice versa)	22	30.1%
Nothing mentioned	2	2.7%
Unclear	6	8.2%

3.3.4 Validation Rigor and Generalization (RQ3)

Validation designs across the corpus were unevenly split between within-event and cross-event or cross-region designs. Twenty of the 73 sources relied solely on a same-event or same-benchmark random split (for example an 80/20 or k-fold split drawn from a single flood event or a single benchmark's native train and test partitions), with no evaluation of transfer to an unseen region, dataset, or event [79, 100, 111, 112]. Twenty-eight sources explicitly incorporated a cross-region, cross-dataset, or cross-event holdout design, and a further 9 combined both a same-event split and a separate cross-region or cross-dataset test within the same study; the validation design of the remaining sources could not be classified into either category with confidence from the reported text. Where cross-region or cross-dataset generalization was tested (37 of 73 sources: the 28 with an explicit cross-region/cross-dataset/cross-event holdout plus the 9 that combined this with a same-event split), the results were mixed but tended to show performance held up reasonably well for open-water-dominated flood scenes and degraded more substantially for vegetated, urban, or terrain-complex scenes. For example, one large benchmark study

explicitly reported that classification accuracy on optical imagery dropped to near-chance levels, about 0.5, over open water in one region, while SAR performance remained close to ceiling in the same region, and the reverse pattern held over vegetated floodplains, with the fused sequence model achieving the best accuracy across all regions when both patterns were pooled [125]. A foundation-model pretraining study found that transfer from a global pretraining corpus to an unseen, domain-shifted regional holdout (Bolivia) retained the large majority of in-distribution performance for the fused configuration, while single-modality baselines degraded more sharply under the same domain shift [119]. Several more recent studies (2025 to 2026) explicitly tested zero-shot transfer to entirely unseen flood events or countries not represented in training, generally reporting a moderate but non- trivial performance drop relative to in-distribution test splits, and in a minority of cases reporting that fusion benefit itself shrank or reversed once evaluated out of distribution, most explicitly for a study that found fusion advantages that were clear within-event became inconsistent when the same architecture was evaluated on a held-out geographic region [113]. The 20 sources noted above as relying solely on a same-event or same-benchmark split did not test cross-region or cross-dataset generalization at all, evaluating exclusively within a single named event or a single regional study area, for example the 2019 Sardoba dam break in Uzbekistan [129], the July 2025 Miyun reservoir flood [116, 137], or the September 2023 Derna dam collapse [139]; the remaining 16 sources' validation design could not be classified with confidence from the reported text and are not counted in either the "tested" or "did not test" totals above. This subset of the literature, while often methodologically careful within its own study area, provides comparatively weak evidence on whether reported fusion gains would hold outside the specific event, sensor geometry, and land-cover conditions under which they were demonstrated, a limitation several of the authors themselves acknowledged as a direction for future work.

Table 14. Aggregated Sentinel-1/Sentinel-2 fusion performance summary by metric type across reviewed studies

Metric Type Reported	Number Studies	Matched S1 vs. S2 vs. Fusion Comparison (%)	Positive Fusion Benefit (%)	Mixed / Context-Dependent Benefit (%)	Cross-Region/ Dataset Generalization Tested (%)
Accuracy / Overall Accuracy (OA)	50	74	96	4	48
Precision / Recall	48	77.1	89.6	8.3	56.3
F1-score / Dice	47	87.2	89.4	8.5	63.8
IoU / mIoU	45	91.1	91.1	6.7	71.1
Kappa Coefficient	17	76.5	100.00	0.0	29.4
AUC / ROC	7	85.7	100.0	0.0	42.9
Visual Assessment Only	1	100.0	100.0	0.0	0.0

3.3.5 Operational Readiness, Cloud-Cover Handling, and Deployment (RQ4)

Sentinel-2 cloud-cover unavailability at the time of a flood event was explicitly addressed, in some technical form, in 69 of the 73 included sources, reflecting near-universal recognition of this constraint across the literature. The most common strategy, appearing across roughly two-thirds of these sources, was to rely on Sentinel-1's cloud-penetrating, all-weather acquisition capability to substitute for or compensate for missing or degraded optical imagery during the flood window, either through simple sensor substitution [125], through cloud-aware dynamic routing or gating that shifts weight toward SAR features specifically in cloud-contaminated pixels [121], or through explicit cloud masking of the optical channel combined with SAR-led classification in masked regions [102, 110]. A second, more recent strategy, appearing mainly in 2024–2026 studies, used generative or reconstruction-based methods to synthesize cloud-free optical imagery directly from SAR, for example GAN-based SAR-to-optical translation [108], diffusion-based SAR-to-electro- optical translation [91], and a cross-modal multitask distillation framework that jointly reconstructs optical imagery and predicts flood extent [140]. A third strategy used temporal compositing or gap-filling across a time series to bridge cloud gaps rather than resolving them within a single acquisition date [143], for example monthly Google Earth Engine compositing [111] or a spatio-temporal Markov random field that explicitly models and fills gaps across a SAR and optical time series [92]. Only 3 sources did not address cloud-cover handling in any explicit technical form, and one further source explicitly stated that no cloud-handling step was implemented because a naturally cloud-free Sentinel-2 acquisition happened to be available for that particular event, which the authors themselves flagged as a favorable but non-generalizable circumstance. Discussion of cloud cover as a constraint on real-world, as opposed to retrospective, deployment was less consistent than the technical handling itself. Many sources discussed cloud cover primarily as a modeling problem to be solved architecturally, with comparatively brief or no discussion of how frequently, and for how long, cloud cover would be expected to block usable Sentinel-2 acquisitions during an actual flood event in a given climate or season; a smaller number of sources discussed this explicitly, for instance noting that heavy cloud cover is essentially inherent to the meteorological conditions that produce flooding in the first place, making continuous optical monitoring impossible during the most operationally critical window [125], or quantifying how much more frequently SAR acquisitions were available than usable optical acquisitions over the same period [125].

Table 15. Effect of joint multimodal training on individual-modality branch performance

Outcome for Single-Modality Branch	Sentinel-1 Branch	Sentinel-2 Branch
Improved under joint training	21	19
Degraded under joint training	0	1
Mixed	1	1
No separate branch architecture used	13	14
Not reported	38	38

Claims of real-time or near-real-time operational deployment, or of deployment readiness in terms of computational cost, inference latency, or infrastructure, were made in 44 of the 73 sources, though the strength of this claim varied considerably. At one end, several sources reported an actual operational or near-operational system, for example a pipeline feeding into a large technology company's operational flood-warning system [76], a cloud-native workflow implemented end-to-end in Google Earth Engine and used for monsoon-season monitoring [111], and a lightweight architecture explicitly benchmarked for millisecond-scale inference latency on consumer-grade GPU hardware as a proxy for deployability [140, 142]. At the other end, a comparable number of sources used deployment-oriented language, phrases such as "rapid mapping," "near real time," or "operational potential," without reporting any latency, throughput, or infrastructure benchmark, and without an operational system actually described. Twenty-four sources made no explicit deployment or operational-readiness claim at all, presenting their work as offline, retrospective research evaluation, and a further 5 could not be classified with confidence. Taken together, these results suggest that, while cloud-cover handling has become a near-universal design consideration in this literature, and while a majority of studies gesture toward operational relevance, comparatively few sources provide the quantitative latency, throughput, or field-validation evidence that would substantiate a genuine, rather than aspirational, claim of real-world deployability, a gap examined further in Section 2.4.

Table 16. Promising Operational Directions

Study	**Innovation**	**Reported evidence**	**Operational significance**
Wang et al. (2023 [97]	Lightweight edge architecture (LSFUnet)	137.97 FPS on Jetson AGX Orin; 0.065M parameters	Enables on-board / edge deployment instead of cloud inference
Ranjan et al. (2026) [147]	Self-supervised foundation-model pretraining (SPEAR)	~1M-parameter embeddings; transfers across 4 external benchmarks	Reduces dependence on large labeled flood datasets
Talreja et al. (2026) [126]	Multitask cross-modal distillation (FLoRA)	IoU 0.57 → 0.71 vs. U-Net-S1 baseline; 22.7 ms/tile	Robust to missing optical modality during cloud cover
Arul et al. (2026) [144]	Cross-region validated production system (MST-DeepLabV3+)	97.1% IoU (Germany) vs. 92.3% IoU (Indonesia)	Closest to operational readiness; code “to be released”

3.4 Discussion

3.4.1 Summary of Evidence

Several limitations affect the scoping review process itself and should be considered when interpreting the synthesis above.

First, because this is a scoping rather than a systematic review, no formal risk-of-bias or methodological quality appraisal was applied to individual sources, consistent with standard PRISMA-ScR guidance. Comparisons that reported a numeric fusion-versus-single-sensor metric on the same event or dataset were treated as standard-confidence evidence for direction-of-effect synthesis. Low-confidence evidence was defined as: (a) preprints or manuscripts under review at the search cutoff, and/or (b) qualitative or visual-only single-sensor-versus-fusion comparisons with no numeric metric. Low-confidence items remain in the corpus and contribute to directional counts but are not pooled into numeric summary statistics. As a result, the review can characterize the direction and approximate range of reported fusion benefit across the literature, but it cannot certify that any individual reported effect size is free from confounding, selective reporting, or optimistic hyperparameter tuning of the proposed fusion model relative to its single-sensor baselines, a concern that applies generically to method papers that both propose and evaluate their own architecture [114, 134, 138].

Second, the underlying data charting relied on natural-language extraction from often lengthy and heterogeneously structured source texts, and a small number of exact duplicate entries were identified and removed during charting. While this process was reviewed for consistency, the free-text nature of many charted fields, particularly the quantitative performance values, key findings, and rationale fields, means some nuance or qualification present in the original sources may not be fully captured in the structured extraction.

Third, this review's information sources and search strategy necessarily reflect a particular window in a very fast-moving field. Given that 40 of the 73 included sources were published in 2025 or 2026 alone, and that several of the most recent sources are preprints or conference papers still under peer review at the time of charting [132, 144], the field is likely to continue evolving quickly, and any synthesis of "current" architectural trends, particularly the shift toward foundation models and state-space architectures noted under RQ1, should be understood as a snapshot rather than a stable end state.

Fourth, English-language restriction and the requirement for peer-reviewed or clearly attributable publication status may have excluded relevant fusion studies published in other languages or disseminated only as government or agency technical reports, which could bias the geographic and institutional representation of the included evidence, particularly for flood events in non- English-speaking regions where local-language technical reporting may be more common than international journal publication.

Fifth, the review's category (b) inclusion criterion, a matched single-sensor- versus-fusion comparison, is inherently dependent on what authors chose to report. Some studies that use fusion as their primary architectural contribution did not report an isolated single-sensor ablation, either because it was outside their intended contribution or because negative or unremarkable single-sensor results were not considered worth reporting; this could inflate the apparent consistency of positive fusion findings in the RQ2

synthesis, since studies where fusion did not help may be systematically less likely to report the comparison that would reveal this, a form of publication and reporting bias that a scoping review of this kind cannot fully correct for, though the ambiguous- case rule to still include narrative-only or visual-only comparisons at low confidence was one attempt to partially mitigate this by capturing weaker positive evidence rather than requiring only strong quantitative results. Finally, the heterogeneity of metrics, datasets, and flood contexts across the 73 included sources, while itself a key finding of this review, means the magnitude comparisons offered above (for example comparing a ten-to-twenty-point IoU gain in one study against a one-to-two-point gain in another) including the pooled mean and median gain figures reported in Table 4, should be read as a descriptive summary of the range and central tendency present in this literature rather than as a formal meta-analytic estimate of a true underlying effect size. The 45 studies pooled for that figure report three different metrics (IoU, Overall Accuracy, and F1) that are not numerically interchangeable, were evaluated on different datasets, land covers, and sample sizes, and were not weighted by study quality, sample size, or precision, all of which a formal meta-analysis would require and which this scoping review design does not attempt to produce.

A small number of included titles sit close to the stated inundation/fusion scope boundary and warrant explicit flagging rather than silent inclusion: refs [110] (probabilistic flood risk mapping, adjacent to the excluded flood susceptibility mapping category), [127] (framed around flood prediction rather than detection/mapping of an already-occurring event), [128] (general remote sensing disaster recognition rather than flood-specific inundation mapping), and [137] (post-disaster flood and building damage assessment, adjacent to the excluded damage assessment category). Each was retained on the basis that its methodology and evaluation still centre on Sentinel-1/Sentinel-2 fusion for mapping flood-affected area at the time of charting.

3.4.2 Conclusions and Implications

Returning to the review's central question, whether Sentinel-1-to-Sentinel-2 fusion measurably improves flood inundation mapping performance over single- sensor approaches, the evidence assembled here supports a qualified yes. Across 73 studies spanning classical machine learning through 2026-era foundation models, 59 reported some form of matched single-sensor-versus-fusion comparison, and fusion outperformed the stronger single-sensor baseline directionally in 65 of the 73 sources overall once qualitative and low-confidence comparisons are included, with 37 of 45 sources (82.2%) among those reporting a directly comparable numeric value confirming the direction quantitatively, and this held broadly across architecture families, geographies, and flood types. Quantitatively, the pooled matched-comparison data (Table 4) put the typical benefit at a mean of +6.2 percentage points and a median of +4.2 percentage points on the primary reported metric, with a right-skewed distribution: roughly a fifth of matched studies (10 of 45) reported gains exceeding 10 points, while a similar fraction (8 of 45) reported no gain or an outright decline. But the qualification matters: the size of that benefit is architecture, modality, and dataset dependent, ranging from decisive (up to +29.9 points IoU) to negligible or negative (as low as -3.0 points) depending on how much headroom existed in the single sensor baselines being compared, and asymmetric branch level benefit, in which one sensor gains disproportionately from joint training while the other does not, is a real and recurring but not universal pattern whose direction depends on fusion mechanism and flood context rather than following a fixed rule [89, 98, 108].

With respect to validation rigor (RQ3), a meaningful share of the literature, roughly a quarter to a third of included sources depending on how borderline cases are classified, has not yet demonstrated that reported fusion gains survive movement to an unseen region, dataset, or event, and the studies that have tested this show that fusion advantage itself can narrow or become inconsistent under such conditions [113]. With respect to operational readiness (RQ4), the field has made cloud-cover handling a nearly universal design consideration, but comparatively few studies back deployment or real-time claims with the latency, throughput, or field-validated evidence that would distinguish a genuinely operational capability from an aspirational one [95, 123, 136, 140, 147].

This "qualified yes" is offered with two corpus-integrity caveats still open at the time of writing, both flagged earlier in this chapter rather than hidden: first, 6 of the 73 records (4 preprints/under-review manuscripts and 2 unclassified by publication type) have not been individually re-adjudicated against the stated peer-reviewed-only eligibility rule (Section 2.3.1); second, 4 further titles sit close to the inundation/fusion scope boundary and have not completed a source-by-source re-adjudication (Section 2.4.1). Neither issue is expected to reverse the qualified-yes conclusion, since the affected records are a small fraction of the 73-source corpus and are not concentrated among the strongest quantitative results, but the precise counts and percentages reported above should be treated as provisional pending that adjudication rather than final.

The implication for the field is that the next phase of research in this space should prioritize demonstrating generalization and deployability at least as much as proposing new fusion architectures. Concretely, this review's findings suggest three areas where future primary research, and future systematic or meta-analytic follow-up work building on this scoping map, would be most valuable: first, systematic cross-region and cross-dataset benchmarking of the fusion architectures already proposed in this corpus, rather than continued proposal of new architectures validated only within-event; second, standardized reporting of matched single-sensor baselines alongside every proposed fusion model, including under data conditions (heavy cloud cover, dense vegetation, urban terrain) specifically chosen to stress-test rather than flatter the fusion advantage; and third, quantified operational benchmarking, inference latency, end-to-end pipeline timing from acquisition to delivered flood map, and field validation against independent ground truth, reported with the same rigor currently applied to segmentation accuracy metrics. Addressing these three gaps would move the evidence base from its current state, a large and rapidly growing set of architecture proposals with generally but not uniformly positive fusion results, toward the kind of operationally actionable evidence base that disaster-response agencies would need to adopt Sentinel-1-to-Sentinel-2 fusion as a standard, rather than experimental, component of flood inundation mapping.

Chapter 4: Flood Mapping Application

4.1 Introduction

4.1.1 Problem Statement

The Copernicus Global Flood Monitoring (GFM) service is a widely used operational satellite-based flood mapping product: it processes Sentinel-1 imagery worldwide and typically delivers flood maps within hours of acquisition, built into a three-algorithm ensemble with contextual exclusion layers rather

than a single threshold [149]. By design, GFM relies exclusively on VV polarisation. Flooded vegetation can produce enhanced co-polarised (VV) backscatter through double-bounce interactions between the water surface and vertical vegetation structure, while cross-polarised VH is more sensitive to volume scattering and depolarisation and is generally less able to register this double-bounce enhancement [150]. The relative value of VV and VH for detecting flooding beneath crop canopies is therefore scene-, canopy-, and geometry-dependent rather than a fixed advantage for either channel, a nuance especially relevant in rice-growing charland environments such as Roumari Upazila. GFM's own developers have flagged additional-channel integration and site-specific evaluation as directions worth pursuing [149]. Independent benchmarks suggest what is at stake: a related JRC model correctly identified only around 40% of flooded pixels globally [151], a separately trained global model has outperformed GFM-family products by a meaningful F1 margin [152], and a recent perspective frames operational SAR flood mapping as a full-stack systems problem in which polarisation is only one of several factors governing real-world performance [153]. What GFM's single-polarisation design costs specifically in Bangladesh's rice-growing charlands has not been empirically measured here.

4.1.2 Research Objective

This chapter benchmarks six standalone Sentinel-1 flood-detection workflows against a named GFM product across four monsoon events (2019, 2020, 2023, and 2024) in Roumari Upazila, Kurigram, using a Google Earth Engine pipeline. The specific objectives are to: (1) map flood extent for each event using six SAR- based detection methods; (2) compare method performance against optical reference data under two validation windows; (3) benchmark all methods against GFM under matched conditions, and test whether any resulting advantage is statistically significant; (4) examine how these comparisons vary from event to event; and (5) describe cropland flood exposure by crop-calendar stage. On this last point, the analysis is deliberately limited in scope: it provides seasonal context for interpreting flood-exposure figures, not a ranking of agricultural damage or vulnerability across events, since year-specific planting dates, flood depth, and inundation duration were not available for this study.

4.1.3 Relevant Literature Context

The literature on SAR-based flood detection converges on a small set of recurring approaches, each exploiting different physical principles and processing strategies. This chapter evaluates six of them side by side: VH change detection, VV change detection, a VV+VH persistence rule (MultiPol), an absolute-VH Otsu threshold, a Random Forest transferred from an external labelled dataset (Stage 1 RF), and a Random Forest trained locally at the study site (Stage 2 RF). A seventh candidate, a VV/VH backscatter-ratio threshold, was also tested during development but is excluded from the results reported here after failing systematically across all four events (Section 3.4.2).

VH change detection computes per-scene differences in VH backscatter against a pre-monsoon baseline composite, applying Otsu adaptive thresholding to the change image. VH is generally more sensitive to volume scattering and depolarisation from vegetation canopies than to the double-bounce enhancement associated with co-polarised backscatter [150], so its effectiveness in vegetated floodplains depends on canopy density and geometry rather than a uniform double-bounce advantage. The method requires a clean baseline composite representing dry-season conditions, typically constructed from multiple acquisitions during the pre-monsoon period (March- April).

VV change detection follows the same procedure using VV co-polarisation backscatter, which is sensitive to specular reflection from open water. It is generally simpler to apply than VH change detection because the VV signal is stronger and more consistent, particularly during early crop stages when canopy cover is limited, conditions relevant to this study area, discussed further in Section 3.4.2.

Multi-polarisation frequency fusion (MultiPol) classifies a pixel as flooded if at least 50% of its pooled VV and VH observations for the event are positive. This threshold is applied over the full set of pooled observations per pixel, not a single VV/VH pair; for the two-scene 2023 and 2024 events this means at least 2 of 4 pooled observations must be positive. The rule aims for broader detection coverage than either channel alone, but its persistence requirement can cost recall by discarding genuine single-channel detections. Otsu thresholding applied to absolute backscatter histograms [154] provides a baseline that does not require a reference composite, at the cost of vulnerability to inter-event soil moisture variation. Threshold-based SAR methods of this kind are already well established across Bangladesh [155], [156], [157]. Random Forest classifiers offer a more flexible approach, incorporating multiple input features (backscatter values, change images, polarisation ratios, topographic indices, and vegetation indices) to learn decision boundaries adapted to local terrain. Where local labelled training data are scarce, the Sen1Floods11 dataset [158] offers a transfer-learning starting point; Google Earth Engine now also supports fully automated, deep-learning-based global flood mapping, recently demonstrated on the 2024 Bangladesh flood itself [159], and Random Forest and related classifiers have shown consistent gains over simple thresholds in vegetated floodplains, a pattern recently synthesised across benchmarking studies [160]. A related line of work maps persistent and seasonal surface water at global scale using adaptive SAR thresholds [161], but targets a different phenomenon (persistent rather than temporary, vegetation-obscured flooding), so it is not used as an accuracy comparison here. Reported flood- detection accuracy also varies systematically by SAR or optical modality and benchmark dataset [162], underscoring that validation design matters as much as the detection method itself. In this study, a two-stage transfer-learning approach (Stage 1 pre-trained on Sen1Floods11, Stage 2 calibrated with local Otsu-derived pseudo-labels) is evaluated as two separate methods, since each is thresholded and assessed independently (Section 3.2.3).

4.2 Methodology

4.2.1 Study Area

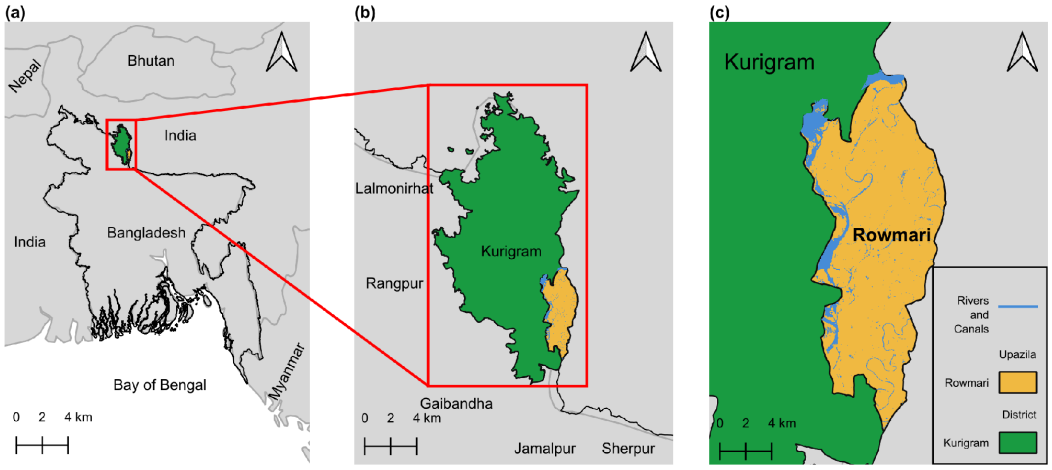


Figure 19. Study Area Map.

The study area encompasses Roumari Upazila, covering approximately 200 km² in Kurigram district, northern Bangladesh, at the intersection of the active Brahmaputra-Jamuna floodplain and the Teesta distributary system [163]. The region has flat terrain (slopes under 5°), seasonally active channels, and morphodynamically unstable charlands reshaped continually by lateral migration and bank erosion [164], driven by the river's roughly 500-800 million tonnes of sediment transported annually [165]. Agricultural land covers approximately 12,117 ha (about 60.5% of the upazila, per a WorldCover-based estimate), predominantly rain-fed T. Aman rice and jute.

Table 17. Hydrological Indicator Matrix (FFWC)

Station	Present WL (mMSL)	Danger Level (mMSL)	RHWL (mMSL)	DLM (m)	HSM (m)	Selected
Singra	9.91	12.2	12.84	2.29	0.64	—
Boalmari	21.6	23.9	24.57	2.3	0.67	✓
Kalmakanda	3.27	6.55	12.6	3.28	6.05	—
Narsingdi	1.25	5.25	6.21	4.00	0.96	—
Thakurgaon	45.76	49.95	50.87	4.19	0.92	—
Nakuagaon	16.98	21.95	25.00	4.97	3.05	—
Chapai-Nawabganj	14.24	20.55	22.39	6.31	1.84	—
Rohanpur	14.43	21.55	23.65	7.12	2.10	—

Hatboalia	6.32	14.05	14.67	7.73	0.62	—

DLM (danger level minus present water level) and HSM (recorded highest water level minus danger level) are descriptive screening indicators computed from FFWC station records on the dates used for site shortlisting. They are not a validated multi-hazard vulnerability index and are time-dependent; they were used only to rank candidate stations for study-area selection, not to infer long-term chronic flood risk rankings for policy use.

Boalmari was selected on a near-minimum Danger Level Margin combined with direct Brahmaputra-Jamuna exposure. Site selection followed a three-stage protocol: (1) BWDB Flood Forecasting and Warning Centre (FFWC) vulnerability maps [166] identified candidate upazilas with direct exposure to major river systems; (2) candidates were cross-checked against the literature; and (3) the Danger Level Margin (DLM, the gap between the official danger level and present water level) and Historical Severity Margin (HSM, the gap between the recorded highest water level and the official danger level) were computed from FFWC station records [166] for all candidates. Table 18 shows Boalmari's DLM (2.3 m) essentially tied with Singra's (2.29 m, the lowest in the shortlist) and its HSM (0.67 m) close to Hatboalia's (0.62 m, the lowest). Boalmari's DLM and HSM are therefore best described as among the lowest in the shortlist rather than the single lowest on either measure; consequently, Boalmari was preferred because both Singra and Hatboalia lack direct Brahmaputra-Jamuna exposure per the FFWC vulnerability maps.
The four events studied had different flood drivers: 2019 and 2020 were Brahmaputra-Jamuna discharge peaks, 2023 reflected elevated late-monsoon river levels, and 2024 was an early-season flash flood, per FFWC event records and bulletins [166]; the underlying station values and dates used for this classification are archived in the project data-provenance log.

4.2.2 Data Sources

All data was processed in Google Earth Engine (GEE). The primary flood signal was Sentinel-1 GRD imagery, IW mode, ascending pass, 10 m resolution, referenced against a pre-monsoon baseline composite built from acquisitions between 1 March and 30 April for each event year. Before the Sentinel-1B failure in December 2021, the constellation had a 6-day repeat cycle, yielding 3-4 flood-window scenes per event; afterward, the single-satellite configuration stretched the revisit interval to 12 days, limiting 2023 and 2024 to two scenes each: an association between satellite status and scene count, not a controlled test of its effect on accuracy.

Table 18. Sentinel-1 Flood-Window Acquisition Manifest

Event	Flood-window S1 scenes	Acquisition dates	Baseline scenes
2019	3 (2× S1A, 1× S1B)	2019-07-19, 2019-07-25, 2019-07-31	5 (1 Mar–30 Apr)
2020	4 (3× S1A, 1× S1B)	2020-07-01, 2020-07-13, 2020-07-19, 2020-07-25	5
2023	2 (S1A)	2023-08-15, 2023-08-27	5
2024	2 (S1A)	2024-06-22, 2024-07-04	4

Optical validation composites came from Sentinel-2 SR Harmonized (10-20 m) and Landsat-8 OLI/TIRS (30 m), using NDWI and MNDWI water indices with per-scene cloud masking via the Sentinel-2 Scene Classification Layer (SCL band). These composites represent surface water over the validation window as a whole, not a pixel-exact match to any single flood-map date (Section 3.5.2).

Supporting datasets: the JRC Global Surface Water v1.4 product [167] masked permanent water bodies; SRTM DEM derivatives (slope, Topographic Wetness Index) fed the Random Forest classifiers; and the ESA WorldCover v200 product (10 m, 2021 epoch) defined the cropland mask used for exposure calculations. Because that mask reflects 2021 land cover, the cropland-exposure figures in Section 3.4.5 should be read as intersections with the mapped 2021 cropland class, not as verified crop occupancy for each event year.

GFM products were accessed through the JRC portal at their native ~20 m resolution [149]. Asset- level metadata (product ID, version, and processing date) came back empty for all four assets used, so the exact GFM product version behind each comparison could not be recovered; this gap is disclosed rather than glossed over. For each event, the GFM comparison scene was selected by visually inspecting the Copernicus EMS GFM archive for the scene showing the most extensive mapped inundation within the event window, then retrieving matching Sentinel-1 and Sentinel-2 acquisitions. Because this scene was picked for maximum apparent extent rather than a fixed offset from event onset, the restricted-window comparison (Section 3.2.4) narrows but does not eliminate the temporal gap between the SAR, optical, and GFM references.

4.2.3 Flood Detection Methods

Six standalone methods were implemented and carried through to the results (Table 20); a seventh, a VV/VH-ratio threshold, was tested but failed systematically across all four events and is excluded (Section 3.4.2). An internal majority-vote ensemble of the six methods was also computed during development but is not reported below, since no claim in this chapter depends on it.

Method 1 (VH Change Detection): Per-scene differences in VH backscatter were computed against the pre-monsoon baseline, with Otsu adaptive thresholding applied to the change image; pixels above the threshold were classified as flooded.

Method 2 (VV Change Detection): The same procedure applied to VV co-polarisation backscatter.

Method 3 (MultiPol): A pixel was flagged flooded if ≥50% of its pooled VV and VH observations for the event were positive. Mean pooled observations per pixel were 6.0 (2019), 8.0 (2020), and 4.0 (2023 and 2024); for the two-scene 2023/2024 events, the rule required at least 2 of the 4 pooled observations to be positive. Persistence discarding genuine detections is a plausible contributor to MultiPol's comparatively lower accuracy (Section 3.4.2), though errors were not broken down by channel and scene, so this explanation is not assumed to hold uniformly.

Method 4 (Otsu Absolute): Otsu thresholding [154] was applied directly to the absolute VH backscatter histogram, without a baseline. This method needs no reference data, at the cost of vulnerability to inter-event soil moisture variation.

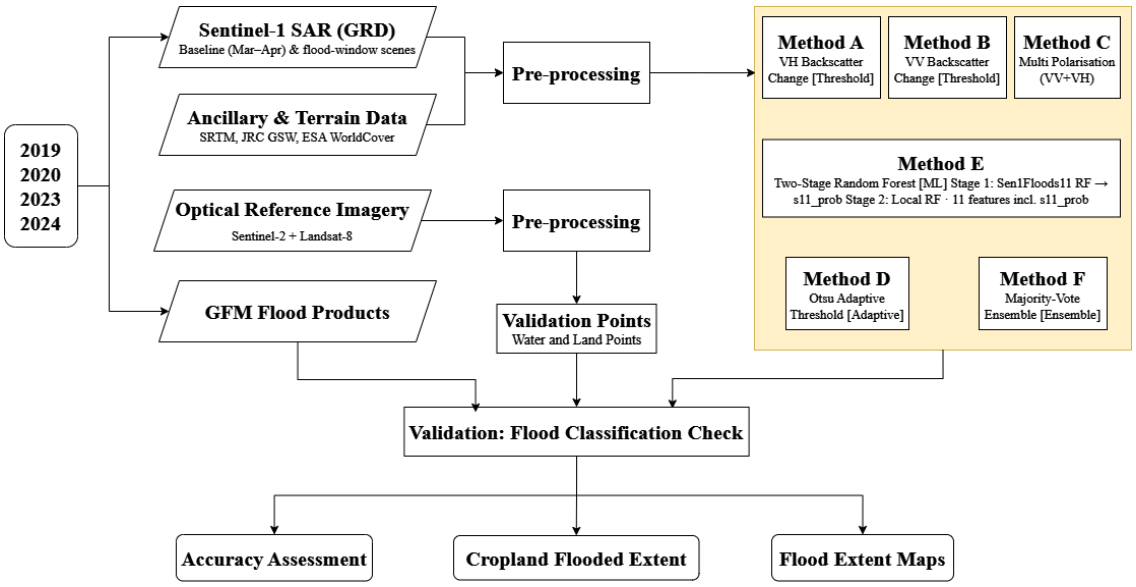


Figure 20. Overview of the Methodological Framework for SAR-Based Flood Detection, Classification, and Validation Across Four Flood Events.

Method 5 (Stage 1 RF): a Random Forest pre-trained on 65,902 labelled pixels from the Sen1Floods11 India subset [158], run in Earth Engine's probability output mode and thresholded at 0.5. Leave-one-chip-out cross-validation across the 68 India training chips (roughly 1,000 evaluation pixels per fold, 63,902 pooled) gave $\kappa = 0.51$, overall accuracy 75.4%, balanced accuracy 75.3%, and F1 74.3%. This source-domain result and the target-domain (Roumari) results reported in Section 3.4 differ in reference protocol, class balance, and sampling design, so the difference between them should not be read as a direct measure of transfer improvement: the Roumari results are additional evidence the model works under this study's target protocol, not proof that transfer was "successful" in a stronger sense.

Method 5 (Stage 2 RF): A second Random Forest trained per event on Otsu-derived flood pseudo-labels, using eleven features: VH/VV backscatter and change, VH/VV ratio, slope, TWI, pre-flood NDVI/NDWI/MNDWI, and the Stage 1 probability layer, thresholded at 0.5. Evaluation points were drawn independently of the training pixels and never used for fitting, threshold selection, feature selection, or model selection. A spatial block cross-validation of Stage 2 against its own Otsu labels (2×3 folds) returned $\kappa = 0.998$, which is expected, since it checks label self-consistency rather than independent accuracy; the claims in Section 3.4 rest on the point-level comparisons against the optical and GFM references, not this internal check.

Implementation note: Classifiers were originally run in Earth Engine's PROBABILITY output mode, documented as the confidence of the predicted class rather than the probability of a specific class, an ambiguous definition for a binary classifier. Both stages were re-implemented in MULTIPROBABILITY mode, which returns an explicit per-class probability array, and the flood-class band was pulled directly. Re-running the full pipeline this way reproduced the original point-level predictions, discordant counts, and bootstrap kappa intervals to floating-point precision, confirming that PROBABILITY mode had in fact matched flood-class probability for all events.

Table 20. Method Inventory

Method	Input(s)	Decision rule	Training/threshold source
VH Change	VH backscatter, pre-monsoon baseline	Otsu threshold on mean VH change image, maximum extent across flood scenes	Unsupervised, per-event Otsu
VV Change	VV backscatter, pre-monsoon baseline	Otsu threshold on mean VV change image, maximum extent across flood scenes	Unsupervised, per-event Otsu
MultiPol	VV and VH per-scene flags	Pixel flagged flooded if ≥50% of pooled VV+VH observations are positive	Unsupervised; inherits per-channel Otsu thresholds
Absolute VH Otsu	Absolute VH backscatter	Otsu threshold on raw VH histogram (no baseline)	Unsupervised, per-event Otsu
Stage 1 RF	Sentinel-1 backscatter features	Random Forest pre-trained on Sen1Floods11 (India subset), probability output thresholded at 0.5	Sen1Floods11 hand labels, 68 chips, 65,902 px
Stage 2 RF	11 features (VH/VV backscatter & change, VH/VV ratio, slope, TWI, pre-flood NDVI/NDWI/MNDWI, Stage 1 probability)	Random Forest trained per event, probability output thresholded at 0.5	Otsu-derived pseudo-labels (see below)

4.2.4 Post-Processing and Validation

Each binary flood map underwent morphological smoothing (opening/closing), exclusion of terrain steeper than 5°, urban texture masking, exclusion of pixels with pre-flood NDVI > 0.4, permanent and seasonal water-body masking via the JRC product, and removal of connected components smaller than 9 pixels (900 m² at 10 m resolution). Random Forest outputs were binarised at a 0.5 probability threshold.

Accuracy was assessed under two temporal windows per event. The extended window (roughly 10 weeks centred on flood peak) maximises optical sample size but introduces temporal latency. The restricted

window aligns with SAR acquisition dates, putting all methods, including GFM, on more comparable footing, and serves as the primary comparison basis throughout; a shared window, however, does not guarantee the optical reference is equally accurate for every method's specific acquisition date.

Table 20. McNemar Comparison Detail for the Best-Performing Local Method Vs. GFM (matched validation points; continuity-corrected asymptotic χ^2 and exact two-sided binomial p)

Event/Window	Method	Local-right/GFM-wrong (b)	Local-wrong/GFM-right (c)	n discordant	χ^2 (df=1)	Exact p
2019 Ext	Stage 2 RF	72	8	80	49.61	$< 1\times10^{-13}$
2019 Res	Stage 1 RF	27	10	37	6.92	0.0076
2020 Ext	Stage 1 RF	49	14	63	18.35	1.1×10^{-5}
2020 Res	Stage 1 RF	34	15	49	6.61	0.0094
2023 Ext	Stage 1 RF	54	8	62	32.66	$< 1\times10^{-8}$
2023 Res	Stage 2 RF	35	17	52	5.56	0.0175
2024 Ext	Stage 1 RF	37	14	51	9.49	0.0018
2024 Res	Stage 1 RF	31	10	41	9.76	0.0015

Note: Paired McNemar tests evaluate whether the best local method and GFM disagree on the same validation points. Bootstrap confidence intervals for κ and for the paired difference Δκ were specified in the analysis plan but are not reported in the thesis tables; the primary inferential claim for the local–GFM gap is therefore the McNemar result, with point-estimate κ values as descriptive summaries.

Stratified random sampling generated 200 validation points per event and window, drawn equally from flooded and non-flooded classes. Because sampling is balanced rather than proportional to landscape flood prevalence, the reported κ, F1, and balanced-accuracy figures describe performance under this design, not unweighted, landscape-wide error rates; this follows the spirit of established good-practice guidance for accuracy and area estimation [168], though the full area-estimation and stratified-weighting framework it describes was not implemented. Points were drawn without enforcing minimum spacing, so reported uncertainty intervals may understate true sampling variance if there is meaningful spatial autocorrelation.

4.2.5 Phenology-Aware Agricultural Exposure Analysis

Each flood map was intersected with the WorldCover 2021 cropland mask to obtain mapped inundated cropland area per method and event. Each event window was then cross-referenced against the local crop calendar for Kurigram District, covering T. Aman, Aus, Boro rice, wheat, and mustard [169]-[174]. Most of these calendar sources describe the wider north-western or Rajshahi region and national-level phenology rather than Roumari or Kurigram specifically, so they provide regional seasonal context rather than site-verified planting dates.

Table 21. Cohen's Kappa (κ) for the best-performing local method vs. GFM, extended (Ext) and restricted (Res) validation windows, across four flood events. Best method is Stage 2 RF (2019 Ext, 2023 Res) or Stage 1 RF (all other cells). Full per-method accuracy metrics (BA, PA, UA, F1) for all six local methods and both windows are intended for the project repository (see Section III.C).

Event	Best method (Ext κ)	GFM (Ext κ)	Best method (Res κ)	GFM (Res κ)	Δκ (Res)
2019	0.680	0.040	0.660	0.490	+0.170
2020	0.600	0.250	0.630	0.440	+0.190
2023	0.760	0.300	0.500	0.320	+0.180
2024	0.640	0.410	0.540	0.330	+0.210

Table 23 summarises calendar overlap only, not confirmed crop occupancy or damage. The 2019 window (19 Jul-02 Aug) overlaps both T. Aman transplanting activity and Aus-harvest activity in the regional calendar; the 2020 window (21 Jun-27 Jul) overlaps the pre-transplanting and Aus-harvest period. These two windows themselves overlap in calendar time (19-27 July), so the regional calendar alone cannot establish that one entire event coincided with transplanting while the other happened strictly beforehand.

Table 22. Flood event windows cross-referenced against the regional Kurigram-area crop calendar. Entries describe calendar overlap, not confirmed crop occupancy or damage.

Event	Window	T. Aman stage (regional calendar)	Aus stage (regional calendar)	Event
2019	19 Jul–02 Aug	Transplanting-onset period	Harvest period	2019
2020	21 Jun–27 Jul	Pre-transplanting period	Harvest period	2020
2023	05 Aug–29 Aug	Later-transplanting period	Post-harvest period	2023
2024	22 Jun–16 Jul	Pre-transplanting period	Pre-harvest period	2024

Note: Year-specific planting-date records, crop-occupancy maps, flood depth, and inundation-duration data were not available for this study. The calendar overlap is therefore reported as seasonal context only: it identifies periods potentially relevant to Aman establishment and Aus harvest, but the available data does not support ranking the four events by crop damage or agricultural vulnerability, and no such ranking is made here.

4.2.6 Sentinel-1 Preprocessing

Sentinel-1 GRD imagery was drawn from the Earth Engine COPERNICUS/S1_GRD collection, which already applies orbit-file correction, thermal noise removal, radiometric calibration to sigma-nought (σ0) backscatter, terrain correction, and storage in dB; no additional instance of these steps was run in the pipeline. For change detection, per-scene backscatter was converted from dB to linear power, divided by the linear pre-monsoon baseline, and the resulting ratio converted back to dB (10·log10 of the linear ratio), a standard decibel-ratio computation, mathematically equivalent to subtracting the two dB values directly, not a second logarithmic transform of already- converted data. A pre-monsoon baseline

composite was constructed from median-compositing acquisitions between 1 March and 30 April for each event year, chosen to minimise contamination from agricultural activity and provide a stable dry-season reference.

4.2.7 Training and Calibration

The Stage 1 Random Forest was pre-trained in probability mode on 65,902 labelled pixels from the Sen1Floods11 India subset [158], providing physical transferability of SAR backscatter relationships across South Asian alluvial environments without requiring local labels; leave-one-chip-out cross-validation on the source data gave $\kappa = 0.51$ (Section 3.2.3), a source-domain figure that is not directly comparable to the target-domain results in Section 3.4. The Stage 2 Random Forest was trained on Otsu-derived flood labels using the eleven features listed in Section 3.2.3. Training samples were generated from the same pre- monsoon baseline used for change detection and were spatially independent of all validation points, to prevent data leakage. Sen1Floods11 contains hand-labelled chips and weak labels derived from Sentinel-1 and Sentinel-2; the exact chip IDs and whether each of the 68 chips is hand-labelled or weak-labelled are recorded in the project Github Repository (https://github.com/mdasifbinkhaled/Flood-Monitoring-Prediction). Class balance and event distribution for this subset are reported in the same repository.

3.2.8 Validation Framework

The dual-window approach addresses a tension in flood-mapping validation: optical composites provide spatially dense reference data but suffer temporal latency from monsoon cloud cover, while the restricted window aligns validation with SAR acquisitions at the cost of optical sample size. The restricted window is the primary comparison basis throughout this chapter, though it narrows rather than eliminates the temporal gap between the SAR, optical, and GFM references, since GFM's own comparison scene was chosen for maximal apparent extent rather than a fixed offset from event onset (Section 3.2.2). By reporting results under both windows, the study characterises the sensitivity of method rankings to validation timing.

4.2.9 Evaluation Metrics

Flooded and non-flooded class proportions are typically imbalanced. F1-score was reported as a complementary metric. Both were computed with 95% confidence intervals from 2,000 paired bootstrap resamples of the shared evaluation points, computed independently per method. These are marginal intervals for each method's own score; their overlap or non-overlap does not test the paired difference between two methods directly and was not used as a significance criterion. Instead, the significance of each best-performing local method's advantage over GFM was assessed directly on the matched point-level predictions using McNemar's test with continuity correction, reported in Section 3.4.3.

4.3 Baseline Models and Comparative Framework

4.3.1 Threshold-Based Methods

Otsu's method [154] automatically determines a threshold from the bimodal histogram of backscatter values, maximising between-class variance. The assumption of a bimodal distribution holds well for open-water flooding in homogeneous terrain but degrades in complex landscapes with mixed pixels,

vegetation, and urban texture. Change-detection methods compute per-scene differences against a reference baseline before thresholding; their main limitation is baseline quality, since seasonal land-cover change, agricultural activity, and atmospheric conditions can introduce false change signals. In this study area, VV Change consistently outperformed VH Change across all four events (Section 3.4.2), consistent with co-polarised backscatter's greater sensitivity to the double-bounce enhancement produced by flooded vegetation and open water [150], though canopy stage, water depth, and local geometry at each event were not isolated to confirm this as the mechanism; this is offered as a physically plausible explanation rather than a demonstrated mechanism, since no controlled VV-only-versus-VV+VH ablation was run. Water-index methods (NDWI, MNDWI) leverage spectral properties of water in optical imagery under clear-sky conditions, but are limited to cloud-free windows, complementary to, rather than a replacement for, SAR-based approaches in a region where monsoon cloud cover frequently exceeds 70% [155]

4.3.2 Machine Learning Classifiers

Random Forest constructs an ensemble of decision trees, each trained on a random subset of the data and feature set, with predictions determined by majority vote. For flood detection this offers: the ability to combine multiple input features beyond raw backscatter (the eleven features used for Stage 2 RF, including topographic indices, temporal composites, vegetation indices, and derived polarisation ratios and change images); feature-importance ranking; comparatively low computational requirements; and robustness to missing values and heterogeneous feature scales.

Stage 1 RF and Stage 2 RF are evaluated in this chapter as two separate, independently thresholded methods rather than a single combined baseline, since each generates its own binary flood map and its own accuracy figures (Section 3.2.3, Section 3.4.2). Together they address the local training-data availability challenge: Stage 1 supplies physical transferability from the

65,902-pixel Sen1Floods11 India subset, and Stage 2 supplies local calibration through Otsu-derived pseudo-labels, using Stage 1's probability output as one of its own input features.

4.3.3 Global Operational Services (GFM)

GFM is among the most prominent openly accessible, fully-automated global flood-mapping products used by international humanitarian organisations, national disaster-management agencies, and researchers [149], [151], [152], which is why it serves as the primary comparison baseline in this chapter, though no independent usage census was consulted to confirm this is the single most-used such product globally. It is built into a three-algorithm ensemble with contextual exclusion layers, using VV polarisation exclusively, a design choice shaped by cost and prior accuracy evidence [149]. GFM's own developers have identified VH-channel integration and site-specific evaluation as directions worth pursuing [149], and its uniform global processing parameters, applied without local calibration, may not be optimal for specific regional conditions such as Roumari's vegetated charlands, where its conservative exclusion masks can, in principle, discard genuine inundation signals.

GFM's conservative design reflects real operational tradeoffs appropriate to its global emergency mandate: cost, false-alarm suppression, and the practical impossibility of a single global service recalibrating itself to every basin's crop calendar and channel morphology [149]. As noted in Section 3.2.2, GFM's asset-level metadata (product ID, version, processing date) was empty for all four assets

used in this study, so the exact GFM product version behind each comparison could not be recovered, a disclosed gap rather than an assumption of a single fixed configuration.

4.4 Results and Analysis

4.4.1 Flood Event Character

Four monsoon flood events were analysed: 2019, 2020, 2023, and 2024, spanning different hydrological magnitudes, SAR data-density regimes, and seasonal timing within the Brahmaputra-Jamuna floodplain at Roumari Upazila. Boalmari station, the selected site, recorded the lowest Danger Level Margin (DLM = 2.3 m) and Historical Severity Margin (HSM = 0.67 m) among all shortlisted candidates (Table 18), consistent with chronic exposure to direct Brahmaputra-Jamuna flooding; as noted in Section 3.2.1, these station-level figures have not been independently re-verified for this revision.

The 2019 and 2020 events were Brahmaputra-Jamuna discharge peaks; the 2023 event reflected elevated late-monsoon river levels; and the 2024 event was an early- season flash flood. Before the December 2021 Sentinel-1B failure, the constellation maintained a 6-day repeat cycle, yielding 3-4 flood scenes for 2019 and 2020; the subsequent single-satellite configuration extended the revisit interval to 12 days, limiting 2023 and 2024 to two SAR scenes each.

4.4.2 Flood Detection Accuracy

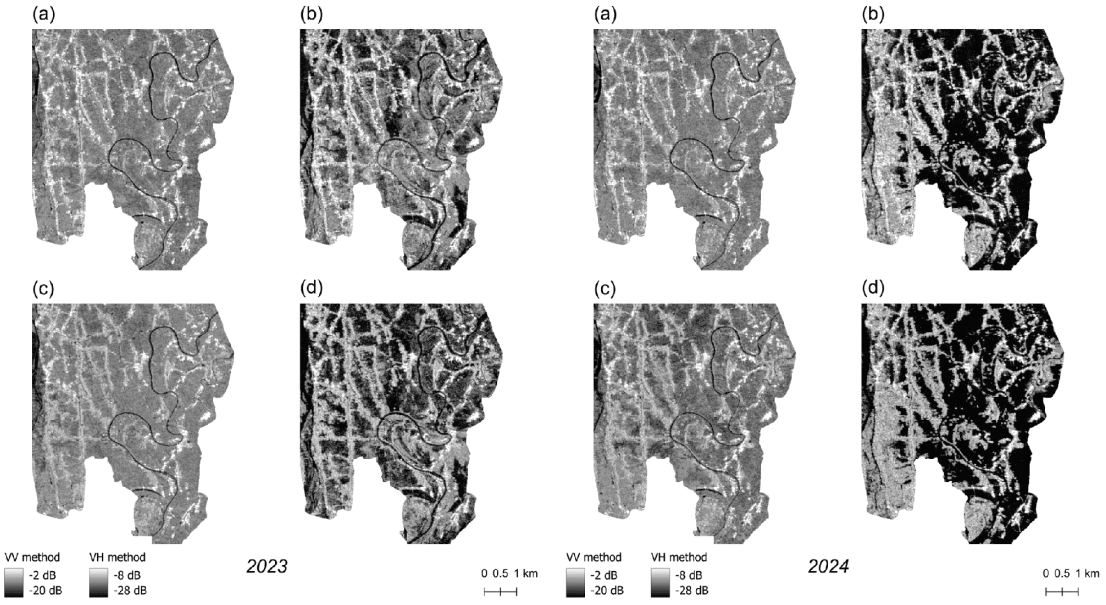


Figure 21 (a). Sentinel-1 SAR backscatter comparison for 2023 and 2024 showing VV baseline (a), VV flood (b), VH baseline (c), and VH flood (d) scenes over the study area.

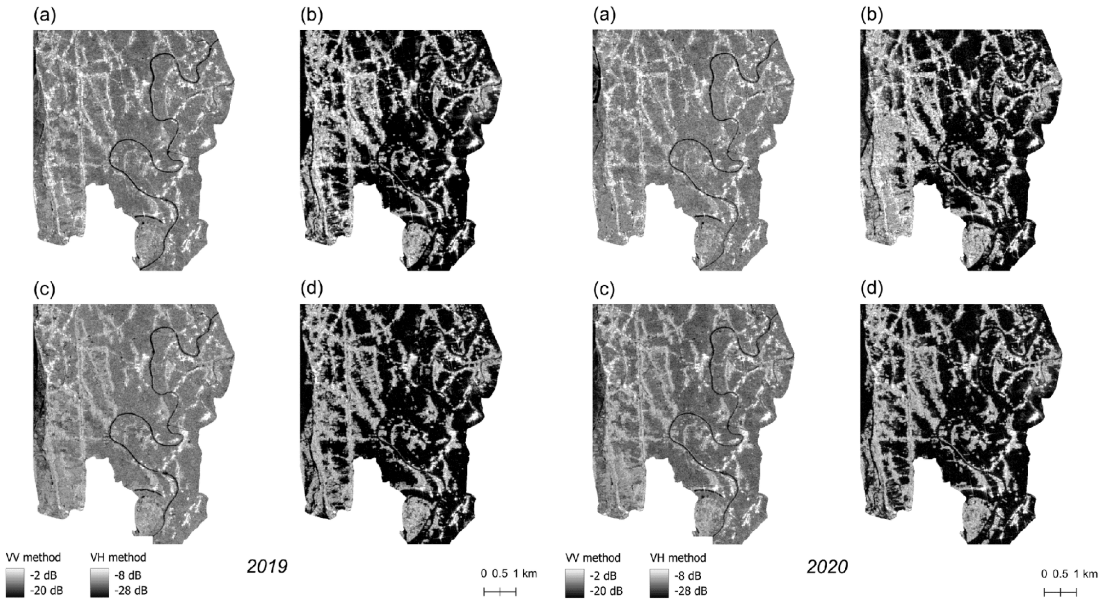


Figure 21 (b). Sentinel-1 SAR backscatter comparison for 2019 and 2020 showing VV baseline (a), VV flood (b), VH baseline (c), and VH flood (d) scenes over the study area.

Under the extended validation window, Random Forest methods outperformed all threshold-based approaches and GFM in every event (Table 22). Stage 2 RF led in 2019 ($\kappa = 0.680$), while Stage 1 RF led in 2020, 2023, and 2024 (κ = 0.600, 0.760, and 0.640, respectively). VV Change consistently outperformed VH Change across all four events. The VV/VH-ratio threshold tested during development failed systematically ($\kappa \leq 0.31$ in every extended-window comparison, as low as 0.01-0.08 under the restricted window) and is excluded from Table 22; an internal majority-vote ensemble across the six reported methods was also computed but is not shown, since no claim below depends on it. Under the restricted window, which aligns all methods, including GFM, on identical SAR acquisition dates, the kappa advantage of the best RF method over GFM was 0.17-0.21 units across all four events (Table 22): 2019 ($\Delta\kappa$ = +0.170), 2020 (+0.190), 2023 (+0.180), and 2024 (+0.210). VV Change alone, requiring no training data, also outperformed GFM in every event under both windows. This indicates the gap between local workflows and GFM cannot be attributed to polarisation availability alone: the comparison changes algorithm, baseline construction, thresholding, temporal aggregation, training, and masking simultaneously between the local methods and GFM, so it does not isolate the marginal contribution of any single ingredient.

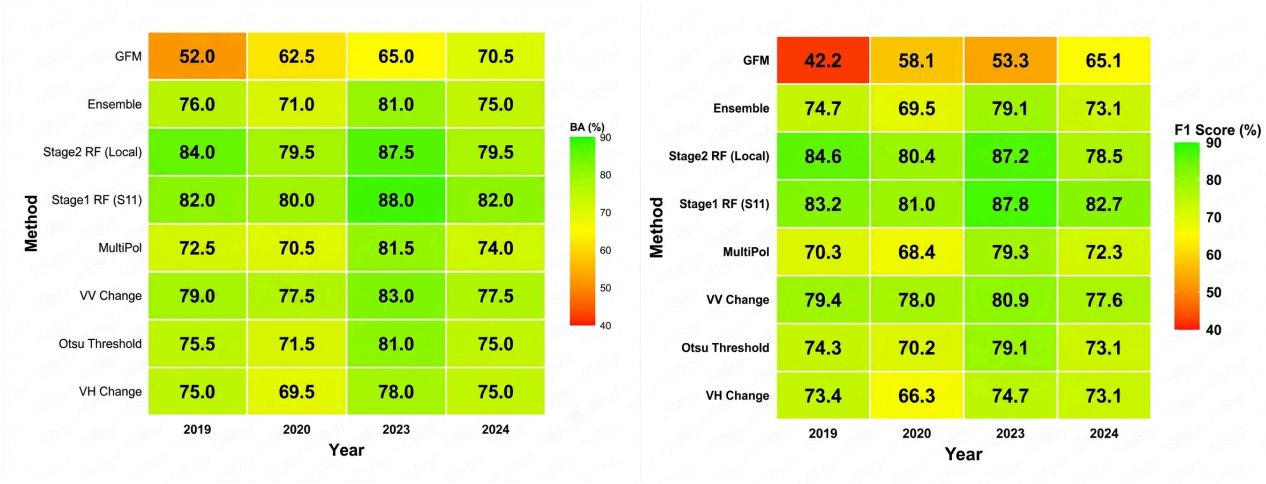


Figure 22. Balanced Accuracy (%) and F1 Score (%) Comparison of different Methods by Year (2019–2024)

4.4.3 Statistical Significance of the Local-GFM Gap

The asymptotic critical value for χ^2 with one degree of freedom at $\alpha = 0.05$ is 3.841; every comparison in Table 21 clears it, and an exact two-sided binomial test on the same discordant counts agrees in direction and significance across all eight event/window comparisons ($p \leq 0.018$ throughout). These counts test paired correctness at the matched evaluation points, not F1 or κ directly; in this study's balanced, unweighted 200-point design the relationship between κ and overall accuracy is straightforward, but that should not be assumed to hold if points are later excluded, reweighted, or class balance changes.

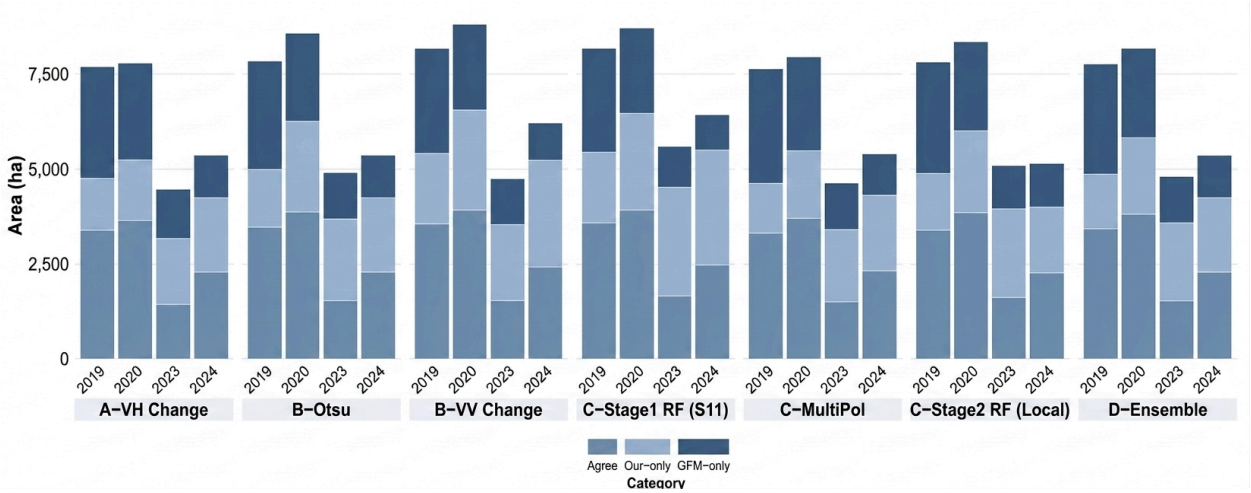


Figure 23. Comparison of flood-mapped area (ha) by detection method across agreement categories (Agree, Our-only, and GFM-only)

Table 23. Spatial Agreement and GFM Capture Rates

Method	2019 Capture	2020 Capture	2023 Capture	2024 Capture	Non-JRC avg.
VH Change	53.8%	59.1%	52.9%	67.6%	48%
VV Change	56.4%	63.6%	56.0%	71.3%	51%
MultiPol	52.6%	60.1%	55.3%	68.5%	49%
Otsu	55.0%	62.8%	56.2%	67.6%	49%
Stage 1 RF	56.8%	63.6%	61.0%	73.2%	49%
Stage 2 RF	53.8%	62.3%	59.1%	66.6%	49%
Ensemble	54.4%	61.8%	55.9%	67.6%	48%

Note: GFM capture rate = agreement area ÷ GFM total mapped area; Non-JRC avg. = share of GFM-only pixels falling outside JRC-classified permanent water, averaged across events.

4.4.4 Spatial Comparison with GFM

SAR-based flood extent broadly agreed with GFM's mapped extent in 2019 and 2020 (SAR-to-GFM ratios 0.80-1.10) but substantially exceeded it in 2023 and 2024 (ratios 1.28-1.84; derived from the mapped-area comparison in Figure 24). Two things line up with this pattern: GFM's exclusion masks appear applied more aggressively in charland zones, where sandy substrates produce backscatter signatures resembling shallow water, and the halving of Sentinel-1 temporal density after the December 2021 Sentinel-1B failure reduced the acquisitions available to GFM in 2023 and 2024. The constellation-density point is treated as a plausible contributor rather than a demonstrated cause, since the four events also differ in driver, magnitude, and season; isolating its effect would need a controlled scene-subsampling experiment, which was not run.

Stage 1 RF achieved the highest GFM capture rate (the share of GFM-flagged flooded area also flagged by the local method) across all four events, 56.8- 73.2% (Table 24). Areas GFM flags that no local method detects ("GFM-only" areas) had 34-64% falling outside permanent water-body classifications, but only 1-5% overlapping JRC-classified wet areas. This pattern is consistent with shallow, possibly vegetation-obscured inundation sitting below the local polarisation-based methods' detection threshold, but it is equally consistent with GFM commission error in charland terrain; telling these apart would require an independent check of a sample of GFM-only disagreement pixels against a higher-resolution reference, which was not done. GFM's mapped extent in these areas is therefore described as smaller, not as an underestimate.

4.5 Discussion

4.5.1 Implications for Flood Monitoring Practitioners

Locally tuned multi-method approaches substantially outperformed the global GFM service in this vegetated charland environment. The 0.17-0.21 kappa-unit accuracy advantage of the best Random Forest

method over GFM, consistent across all four events under the temporally matched restricted window and confirmed as statistically significant in all eight event/window comparisons by paired McNemar tests (Section 3.4.3), represents a structural difference that could plausibly affect damage assessment, agricultural-loss estimation, and emergency- response allocation. These comparisons describe differences between complete workflows, however; on their own they do not isolate polarisation, local thresholding, temporal aggregation, training, or masking as the specific cause of GFM's lower scores, and a controlled ablation would be needed to make that stronger claim. The local approach also requires computational infrastructure, domain expertise, and locally calibrated training data that may not be available in resource-constrained operational settings.

Absolute agreement is highly sensitive to the validation window. For the 2019 event, GFM Cohen's κ changed from 0.040 (extended window) to 0.490 (restricted window). The ranking of the best local method versus GFM remained in the same direction under both windows, but absolute scores and the size of the gap should not be read as window-invariant. Restricted-window results are the primary operational comparison; extended-window results are reported as a sensitivity check.

A pragmatic recommendation is the development of VH-augmented GFM variants that incorporate cross-polarisation data for vegetated environments, reducing the accuracy gap while retaining GFM's global processing infrastructure, potentially incorporating recent lightweight deep-learning architectures shown to perform well on global SAR flood benchmarks [175]. The phenology-aware exposure analysis reinforces that flood extent alone is an insufficient proxy for agricultural- damage assessment, supporting the integration of crop-phenology data into operational flood monitoring, while stopping short of ranking the study's own four events by severity (Section 3.4.5).

Exporting the full point-level evaluation data and computing a paired bootstrap uncertainty estimate for Δκ itself remain open items and are the priority next steps before this approach could be recommended operationally. Future work should evaluate additional events from the Sentinel-1C and Sentinel-1D era (with Sentinel-1D reaching full operational status on 1 May 2026), using actual scene availability over the study area, and should consider a controlled scene- subsampling experiment to separate the effect of observation density from other differences between flood events, before extending this local-calibration approach to the other high-vulnerability charland stations identified during site selection.

4.5.2 Limitations

Optical validation composites carry residual temporal latency from monsoon cloud cover; the restricted window reduces but does not eliminate this, and the latency may not affect all methods and GFM equally, since GFM's own comparison scene was chosen for maximal apparent extent rather than a fixed offset from event onset. More broadly, this study pairs a global product built for conservative, planetary-scale reliability against locally tuned workflows (a real-world operational contrast rather than a controlled single-variable experiment), so its conclusions describe these specific compared workflows, not the isolated effect of polarisation, algorithm, or training.

Post-processing excluded pixels with pre-flood NDVI > 0.4 to suppress false positives over dense vegetation; in 2019 this removed 27-39% of raw flood- classified pixels across methods, disproportionately over actively cropped land. A smaller, filtered map is a subset of the unfiltered one, not automatically a lower bound on true flooding, since some retained unmasked area could still be a false detection and the true flood status of excluded vegetated pixels cannot be established with the evidence available; the direction of any resulting bias remains unresolved.

The 0.5 probability threshold used to binarise RF outputs was fixed in advance and not tuned per event. A sensitivity sweep over t = 0.3-0.7 showed κ varying by no more than 0.01-0.02 units around the production value in most events; in 2020 and 2024 the swept accuracy was marginally higher at other thresholds (2020: κ = 0.550 at t = 0.3/0.6 vs. 0.540 at t = 0.5; 2024: κ = 0.620-0.630 at t = 0.6/0.7 vs. 0.610 at t = 0.5), while 2019 and 2023 were flat or peaked right at t = 0.5. This indicates the reported results are not an artefact of favourable post-hoc threshold selection, though it does not by itself validate the underlying probability-mode semantics discussed in Section 3.2.3. Validation points were drawn by stratified random sampling without enforcing minimum spacing, so the reported confidence intervals may understate true sampling uncertainty if flood and non-flood patches are spatially autocorrelated beyond the 10 m sampling grid. The static ESA WorldCover 2021 cropland mask does not track annual char migration or crop-cycle change, and the uncertainty this introduces into the absolute cropland-exposure estimates has not been quantified. GFM's asset-level metadata (product ID, version, processing date) was unavailable for all four assets used, so the exact GFM product version behind each comparison could not be recovered. Finally, this chapter's findings are limited to a single study area and four flood events, and Stage 2 RF's training relies on Otsu-derived pseudo-labels rather than field-verified optical reference labels.

Chapter 5: Ethical Considerations

5.1 Introduction

This report does not involve human or animal subjects, and no primary data were collected from individuals. Ethical considerations instead center on three areas: research and reporting integrity in the two evidence syntheses (Chapters 2–3), data provenance and licensing across all three studies, and the humanitarian and social value of applying ML-based flood detection in Roumari Upazila, Bangladesh (Chapter 4).

5.2 Research and Evidence-Synthesis Integrity

Both reviews applied recognised safeguards to keep the synthesis faithful to the underlying evidence. GRADE certainty assessment was applied throughout, reported as LOW across all five learned-index outcomes, and PRISMA 2020 / PRISMA-ScR checklists guided study selection. Both reviews also name the specific factors shaping evidence reliability in their fields, author-led self-evaluation in the learned-index corpus, and reliance on custom, unreleased datasets in the flood-mapping corpus, consistent with the report's authenticity declaration.

5.3 Data Provenance, Licensing, and Reproducibility

Chapter 4 relies entirely on openly licensed data: Sentinel-1 SAR and Sentinel-2 multispectral imagery under the Copernicus open and free data policy, and the Copernicus Global Flood Monitoring (GFM) service as a public benchmark, keeping the comparison reproducible by third parties. All 65 learned-index and 73 flood-mapping studies are cited to their original authors following international citation standards. Chapter 3 also notes that only a minority of reviewed flood-mapping studies released code, a field-wide reproducibility gap flagged transparently here, and one this report's own methods do not replicate.

5.4 Humanitarian and Social Implications

Chapter 4's findings bear directly on disaster response in an active floodplain community. The report finds that the locally calibrated Random Forest model outperforms GFM by 0.17–0.21 Cohen's Kappa in vegetated charland terrain, a result with real value for improving flood-response accuracy, and one this report communicates clearly to stakeholders rather than presenting only as a performance statistic. The crop-calendar overlay, showing that the 2019 event overlapped key transplanting and harvest windows despite not being the largest by extent, usefully broadens disaster-response prioritisation beyond extent alone toward timing-sensitive impacts on smallholder farmers.

GFM's strength is its global, open-access availability, while the locally calibrated model's accuracy advantage depends on local training data and expertise. Recognising this points toward a constructive institutional role, most plausibly for agencies such as the Bangladesh Water Development Board's Flood Forecasting and Warning Centre, in extending these accuracy gains to communities that could not otherwise produce them on their own.

5.5 Environmental and Sustainability Considerations

Sustainability was an explicit design consideration in this report. The Random Forest classifier used in Chapter 4 was chosen in part for its low computational footprint relative to the transformer-based, diffusion-regularised, and self-supervised architectures surveyed in Chapter 3, achieving strong accuracy in this setting without the training and inference energy overhead of larger deep-learning approaches. The report's reliance on the shared, publicly funded Copernicus Sentinel constellation (Section 5.3) similarly avoids duplicated data-acquisition effort across research groups. Fully quantifying the energy savings of either choice would require training-log and lifecycle data outside this report's scope, and is noted as a promising direction for future work.

5.6 Scope, Validation, and Responsible Use

This report's findings, including the LOW certainty ratings in Chapter 1 and the accuracy comparison in Chapter 3, are scoped to the specific study area, four flood events, and evidence base examined, and are offered as a rigorously evidenced contribution rather than a final verdict. Readers, including database engineers and disaster-response practitioners, are encouraged to treat reported effect sizes as indicative and to pair them with context-specific validation before operational adoption, in keeping with the evidentiary transparency this report models throughout.

5.7 Addressing Ethical Issues in Deployment

Because the local-Random-Forest-versus-GFM comparison has direct bearing on disaster-response decisions, any future operational adoption should be accompanied by clear communication to end users, such as disaster-management officials, about validation conditions and known failure modes (Section 4.5.2), so the model's demonstrated accuracy advantage is understood in context. Flood-extent data of the kind produced in Chapter 4 could in principle also inform other, non-humanitarian applications, such as land-use or insurance decisions; none were identified during this project, and the ethical review conducted

in this chapter, covering the report's own research conduct, provides a sound foundation for any future downstream application to build on responsibly.

Bibliography

1. Kraska, T., Beutel, A., Chi, E. H., Dean, J., & Polyzotis, N. (2018, May). The case for learned index structures. In Proceedings of the 2018 international conference on management of data (pp. 489-504).
2. Tang, C., Wang, Y., Dong, Z., Hu, G., Wang, Z., Wang, M., & Chen, H. (2020, February). XIndex: a scalable learned index for multicore data storage. In Proceedings of the 25th ACM SIGPLAN symposium on principles and practice of parallel programming (pp. 308-320).
3. Li, P., Hua, Y., Jia, J., & Zuo, P. (2021). FINEdex: a fine-grained learned index scheme for scalable and concurrent memory systems. Proceedings of the VLDB Endowment, 15(2), 321-334.
4. Mo, Y., & Hua, Y. (2025, March). LOFT: A Lock-free and Adaptive Learned Index with High Scalability for Dynamic Workloads. In Proceedings of the Twentieth European Conference on Computer Systems (pp. 475-491).
5. Bhardwaj, G., Chatterjee, B., Sharma, A., Peri, S., & Nayak, S. (2024, August). Kanva: A Lock-free Learned Search Data Structure. In Proceedings of the 53rd International Conference on Parallel Processing (pp. 252-261).
6. Wu, J., Zhang, Y., Chen, S., Wang, J., Chen, Y., & Xing, C. (2021). Updatable learned index with precise positions. Proceedings of the VLDB Endowment, 14(8), 1276-1288.
7. Zhou, W., & Yang, S. (2024, May). SLIPP: a space-efficient learned index for string keys. In Proceedings of the 2024 6th International Conference on Big-data Service and Intelligent Computation (pp. 69-77).
8. Song, Y., Cai, M., Ye, B., & Cheng, G. (2024, November). SLIN: A CPU-efficient, Hybrid Tree and Learned Index for String Data. In Proceedings of the 2024 8th International Conference on Computing and Data Analysis (pp. 52-58).
9. Wang, Z., Feng, J., Dun, L., Bao, Z., & Du, C. (2025). SwiftKV: A Metadata Indexing Scheme Integrating LSM-Tree and Learned Index for Distributed KV Stores. Future Internet, 17(9), 398.
10. Heidari, A., Ahmadi, A., & Zhang, W. (2025). DobLIX: A Dual-Objective Learned Index for Log-Structured Merge Trees. Proceedings of the VLDB Endowment, 18(11), 3965-3978.
11. Wang, Y., Tang, C., Wang, Z., & Chen, H. (2020, August). Sindex: a scalable learned index for string keys. In Proceedings of the 11th ACM SIGOPS Asia-Pacific Workshop on Systems (pp. 17-24).
12. Spector, B., Kipf, A., Vaidya, K., Wang, C., Minhas, U. F., & Kraska, T. (2021). Bounding the last mile: Efficient learned string indexing. arXiv preprint arXiv:2111.14905.
13. Kim, M., Hwang, J., Heo, G., Cho, S., Mahajan, D., & Park, J. (2024). Accelerating string-key learned index structures via memoization-based incremental training. arXiv preprint arXiv:2403.11472.
14. Ferragina, P., Frasca, M., Marinò, G. C., & Vinciguerra, G. (2023). On nonlinear learned string indexing. IEEE Access, 11, 74021-74034.
15. Lu, B., Ding, J., Lo, E., Minhas, U. F., & Wang, T. (2021). APEX: a high-performance learned index on persistent memory. Proceedings of the VLDB Endowment, 15(3), 597-610.

16. Zhang, Z., Chu, Z., Jin, P., Luo, Y., Xie, X., Wan, S., ... & Rudoff, A. (2022). PLIN: A persistent learned index for non-volatile memory with high performance and instant recovery. Proceedings of the VLDB Endowment, 16(2), 243-255.
17. Dai, Y., Xu, Y., Ganesan, A., Alagappan, R., Kroth, B., Arpaci-Dusseau, A., & Arpaci-Dusseau, R. (2020). From WiscKey to Bourbon: A Learned Index for Log-Structured Merge Trees. In 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20) (pp. 155-171).
18. Wang, W., & Du, D. H. C. (2024). LearnedKV: Integrating LSM and Learned Index for Superior Performance on Storage. arXiv preprint arXiv:2406.18892.
19. Liu, Q., Han, S., Qi, Y., Peng, J., Li, J., Lin, L., & Chen, L. (2024). Why are learned indexes so effective but sometimes ineffective?. arXiv preprint arXiv:2410.00846.
20. Sun, Z., Zhou, X., & Li, G. (2023). Learned index: A comprehensive experimental evaluation. Proceedings of the VLDB Endowment, 16(8), 1992–2004.
21. Wongkham, C., Lu, B., Liu, C., Zhong, Z., Lo, E., & Wang, T. (2022). Are updatable learned indexes ready?. Proceedings of the VLDB Endowment, 15(11), 3004-3017.
22. Maltry, M., & Dittrich, J. (2022). A critical analysis of recursive model indexes. Proceedings of the VLDB Endowment, 15(5), 1079–1091.
23. Ferragina, P., & Vinciguerra, G. (2020). The PGM-index: a fully-dynamic compressed learned index with provable worst-case bounds. Proceedings of the VLDB Endowment, 13(8), 1162-1175.
24. Galakatos, A., Markovitch, M., Binnig, C., Fonseca, R., & Kraska, T. (2019, June). Fitting-tree: A data-aware index structure. In Proceedings of the 2019 international conference on management of data (pp. 1189-1206).
25. Ding, J., Minhas, U. F., Yu, J., Wang, C., Do, J., Li, Y., ... & Kraska, T. (2020, June). ALEX: an updatable adaptive learned index. In Proceedings of the 2020 ACM SIGMOD international conference on management of data (pp. 969-984).
26. Yang, Y., & Chen, S. (2024). LITS: An Optimized Learned Index for Strings (An Extended Version). arXiv preprint arXiv:2407.11556.
27. Abu-Libdeh, H., Altınbüken, D., Beutel, A., Chi, E. H., Doshi, L., Kraska, T., ... & Olston, C. (2020). Learned indexes for a google-scale disk-based database. arXiv preprint arXiv:2012.12501.
28. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj, 372.
29. Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Version 2.3, EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.
30. Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., ... & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology, 64(4), 383-394.
31. Wang, H., Wang, X., Ge, J., Chai, Y., Liang, L., & Yi, P. (2025). High Performance or Low Memory? An Updatable Learned Index Framework for Time-Space Tradeoff. Proceedings of the ACM on Management of Data, 3(6), 1-26.

32. Zhang, Z., Jin, P. Q., Wang, X. L., Lv, Y. Q., Wan, S. H., & Xie, X. K. (2021). COLIN: A cache-conscious dynamic learned index with high read/write performance. Journal of Computer Science and Technology, 36(4), 721-740.
33. Wu, S., Cui, Y., Yu, J., Sun, X., Kuo, T. W., & Xue, C. J. (2022). NFL: robust learned index via distribution transformation. Proceedings of the VLDB Endowment, 15(10), 2188-2200.
34. Wang, H., Wang, X., Ge, J., Liang, L., & Yi, P. (2025, April). ShapeShifter: Workload-Aware
35. Adaptive Evolving Index Structures Based on Learned Models. In Proceedings of the ACM on Web Conference 2025 (pp. 4111-4123).
36. Liang, L., Yang, G., Hadian, A., Croquevielle, L. A., & Heinis, T. (2024). SWIX: A memory-efficient sliding window learned index. Proceedings of the ACM on Management of Data, 2(1), 1-26.
37. Amarasinghe, K., Choudhury, F., Qi, J., & Bailey, J. (2024). Learned Indexes with Distribution Smoothing via Virtual Points. arXiv preprint arXiv:2408.06134.
38. Wang, W., & Du, D. H. C. (2025, September). TieredKV: A Tiered LSM-Learned Index Design for Superior Performance on Storage. In Proceedings of the 18th ACM International Systems and Storage Conference (pp. 107-121).
39. Wang, Y., Yuan, J., Wu, S., Liu, H., Chen, J., Ma, C., & Qin, J. (2024, May). Leaderkv: Improving read performance of kv stores via learned index and decoupled kv table. In 2024 IEEE 40th International Conference on Data Engineering (ICDE) (pp. 29-41). IEEE.
40. Tan, H., Li, H., Wang, X., Huang, W., Huang, R., Deng, S., & Wu, J. (2024, December). HL-LSM: A LSM-Tree Combined with Read Hotness and Learned Index. In 2024 IEEE International Conference on Big Data (BigData) (pp. 3992-3999). IEEE.
41. Liu, M., Gu, J., & Zhao, T. (2024, November). Lckv: Learner-cleaner optimized adaptive key-value separated lsm-tree store. In 2024 IEEE 42nd International Conference on Computer Design (ICCD) (pp. 479-482). IEEE.
42. Zhang, Y., Xiong, X., & Balmau, O. (2022, April). TONE: cutting tail-latency in learned indexes. In Proceedings of the Workshop on Challenges and Opportunities of Efficient and Performant Storage Systems (pp. 16-23).
43. Chatterjee, S., Pekala, M. F., Kruglyak, L., & Idreos, S. (2024). Limousine: Blending learned and classical indexes to self-design larger-than-memory cloud storage engines. Proceedings of the ACM on Management of Data, 2(1), 1-28.
44. Maharjan, S., Zhao, S., Zhong, C., & Jiang, S. (2023, December). From leanstore to learnedstore: Using a learned index to improve database index search. In 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS) (pp. 162-169). IEEE.
45. Zhu, R., Wang, H., Xia, S., & Zheng, B. (2025). Learned index for non-key queries. Knowledge and Information Systems, 67(1), 497-519.
46. Zhu, W., Shen, Z., Wei, Q., Chen, R., Yao, X., Yu, D., ... & Shao, Z. (2025). {HiDPU}: A {DPU-Oriented} Hybrid Indexing Scheme for Disaggregated Storage Systems. In 23rd USENIX Conference on File and Storage Technologies (FAST 25) (pp. 271-285).
47. Cui, L., Yang, K., Li, Y., Wang, G., & Liu, X. (2025). Towards optimizing learned index for high performance, memory efficiency and NUMA awareness. ACM Transactions on Architecture and Code Optimization, 22(3), 1-26.

48. Zhang, X., Liang, L., Ailamaki, A., & Xu, J. (2026). HIRE: A Hybrid Learned Index for Robust and Efficient Performance under Mixed Workloads. Proceedings of the ACM on Management of Data, 4(1 (SIGMOD), 1-25.
49. Dong, H., Wang, W., Liu, C., & Du, D. (2025). BLI: A High-performance Bucket-based Learned Index with Concurrency Support. arXiv preprint arXiv:2502.10597.
50. Heidari, A., Ahmadi, A., & Zhang, W. (2024, August). Uplif: An updatable self-tuning learned index framework. In International Database Engineered Applications Symposium (pp. 345-362). Cham: Springer Nature Switzerland.
51. Ge, J., Zhang, H., Shi, B., Luo, Y., Guo, Y., Chai, Y., ... & Pan, A. (2023). SALI: A scalable adaptive learned index framework based on probability models. Proceedings of the ACM on Management of Data, 1(4), 1-25.
52. Guo, N., Wang, Y., Sun, W., Gu, Y., Qi, J., Liu, Z., ... & Yu, G. (2024, May). Chameleon: Towards update-efficient learned indexing for locally skewed data. In 2024 IEEE 40th international conference on data engineering (ICDE) (pp. 4316-4328). IEEE.
53. Lan, H., Bao, Z., Culpepper, J. S., Borovica-Gajic, R., & Dong, Y. (2024, May). A fully on-disk updatable learned index. In 2024 IEEE 40th International Conference on Data Engineering (ICDE) (pp. 4856-4869). IEEE.
54. Zhang, R., Huang, Y., Liang, S., Sun, S., Ma, S., Huan, C., ... & Wu, J. (2024, August). Revisiting learned index with byte-addressable persistent storage. In Proceedings of the 53rd International Conference on Parallel Processing (pp. 929-938).
55. Li, P., Hua, Y., Jia, J., & Zuo, P. (2023). A Fast Learned Key-Value Store for Concurrent and Distributed Systems. IEEE Transactions on Knowledge and Data Engineering, 36(6), 2301-2315.
56. Li, P., Lu, H., Zhu, R., Ding, B., Yang, L., & Pan, G. (2023). DILI: A distribution-driven learned index. Proceedings of the VLDB Endowment, 16(9), 2212-2224.
57. Heidari, A., Ahmad, A., Zhang, W., & Xiong, Y. (2025). Sig2Model: A Boosting-Driven Model for Updatable Learned Indexes. arXiv preprint arXiv:2509.20781.
58. Liu, J., Zhang, F., Lu, L., Qi, C., Guo, X., Deng, D., ... & Du, X. (2024). G-learned index: Enabling efficient learned index on GPU. IEEE Transactions on Parallel and Distributed Systems, 35(6), 950-967.
59. Liu, L., Li, C., Zhang, Z., Liu, Y., Zhou, K., & Zhang, J. (2022, August). A data-aware learned index scheme for efficient writes. In Proceedings of the 51st International Conference on Parallel Processing (pp. 1-11).
60. Kipf, A., Marcus, R., van Renen, A., Stoian, M., Kemper, A., Kraska, T., & Neumann, T. (2020, June). RadixSpline: a single-pass learned index. In Proceedings of the third international workshop on exploiting artificial intelligence techniques for data management (pp. 1-5).
61. Ma, C., Yu, X., Li, Y., Meng, X., & Maoliniyazi, A. (2022). FILM: a fully learned index for larger-than-memory databases. Proceedings of the VLDB Endowment, 16(3), 561-573.
62. Mishra, M., & Singhal, R. (2021, June). RUSLI: real-time updatable spline learned index. In Proceedings of the Fourth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management (pp. 1-8).
63. Qiao, P., Zhang, Z., Li, Y., Yuan, Y., Wang, S., Wang, G., & Yu, J. X. (2024). AStore: Uniformed Adaptive Learned Index and Cache for RDMA-Enabled Key-Value Store. IEEE Transactions on Knowledge and Data Engineering, 36(7), 2877-2894.

64. Ramadhan, A. R., Choi, M. G., Chung, Y., & Choi, J. (2023). An Empirical Study of Segmented Linear Regression Search in LevelDB. Electronics, 12(4), 1018.
65. Song, S., Jin, P., Chu, Z., Luo, Y., & Wan, S. (2023, December). Lifm: A persistent learned index for flash memory. In 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS) (pp. 282-287). IEEE.
66. Tong, A., Hua, Y., & Chen, M. (2025, June). DALdex: A DPU-Accelerated Persistent Learned Index via Incremental Learning. In Proceedings of the 39th ACM International Conference on Supercomputing (pp. 535-549).
67. Wei, X., Chen, R., Chen, H., & Zang, B. (2021). Xstore: Fast rdma-based ordered key-value store using remote learned cache. ACM Transactions on Storage (TOS), 17(3), 1-32.
68. Wang, Z., Chen, H., Wang, Y., Tang, C., & Wang, H. (2022). The concurrent learned indexes for multicore data storage. ACM Transactions on Storage (TOS), 18(1), 1-35.
69. Wang, Z., Ding, C., Song, F., Lu, K., Wan, J., Tan, Z., ... & Li, G. (2024). Wipe: a write-optimized learned index for persistent memory. ACM Transactions on Architecture and Code Optimization, 21(2), 1-25.
70. Yang, Y., Wang, F., Lei, M., Zhang, P., & Feng, D. (2025, May). ALT-Index: A Hybrid Learned Index for Concurrent Memory Database Systems. In 2025 IEEE 41st International Conference on Data Engineering (ICDE) (pp. 86-98). IEEE.
71. Yang, J., Yoon, H., Yun, G., Noh, S. H., & Choi, Y. R. (2025). A Dynamic Characteristic Aware Index Structure Optimized for Real-world Datasets. ACM Transactions on Storage, 21(2), 1-32.
72. Yu, Z., Jin, P., Chu, Z., Luo, Y., Wang, X., & Wan, S. (2023, December). DTtree: A Novel Read/Write-Optimized Learned Index for Database Systems. In 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS) (pp. 308-313). IEEE.
73. Zhang, J., Su, K., & Zhang, H. (2024). Making in-memory learned indexes efficient on disk. Proceedings of the ACM on Management of Data, 2(3), 1-26.
74. Luo, Y., Xie, M., Tong, Y., Jiang, S., & Chai, Y. (2025). Understanding Robustness Issues of Updatable Learned Indexes:[Experiments & Analysis]. Proceedings of the ACM on Management of Data, 3(4), 1-25.
75. Al-Mamun, A., Wu, H., He, Q., Wang, J., & Aref, W. G. (2025). A survey of learned indexes for the multi-dimensional space. ACM Computing Surveys, 58(4), 1-37
76. Sghaier, M. O., Foucher, S., & Landry, T. (2019, July). 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* (pp. 122-125). IEEE.
77. Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., & Datcu, M. (2020). Flood detection in time series of optical and sar images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*(B2), 1343-1346.
78. Kocaman, S., Tavus, B., Nefeslioglu, H. A., Karakas, G., & Gokceoglu, C. (2020). Evaluation of floods and landslides triggered by a meteorological catastrophe (Ordu, Turkey, August 2018) using optical and radar data. *Geofluids*, *2020*(1), 8830661.
79. Tavus, B. E. S. T. E., Kocaman, S., Nefeslioglu, H. A., & Gokceoglu, C. (2020). A fusion approach for flood mapping using sentinel-1 and sentinel-2 datasets. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 641-648.

80. Landuyt, L., Verhoest, N. E., & Van Coillie, F. M. (2020). Flood mapping in vegetated areas using an unsupervised clustering approach on sentinel-1 and-2 imagery. *Remote Sensing*, *12*(21), 3611.
81. Konapala, G., Kumar, S. V., & Ahmad, S. K. (2021). Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *180*, 163-173.
82. Bai, Y., Wu, W., Yang, Z., Yu, J., Zhao, B., Liu, X., ... & 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*(11), 2220.
83. Gašparović, M., & Klobučar, D. (2021). Mapping floods in lowland forest using sentinel-1 and sentinel-2 data and an object-based approach. *Forests*, *12*(5), 553.
84. Yang, Z., Wei, J., Deng, J., Gao, Y., Zhao, S., & He, Z. (2021). Mapping outburst floods using a collaborative learning method based on temporally dense optical and SAR data: A case study with the Baige landslide dam on the Jinsha River, Tibet. *Remote Sensing*, *13*(11), 2205.
85. Soria-Ruiz, J., Fernandez-Ordonez, Y. M., Ambrosio-Ambrosio, J. P., Escalona-Maurice, M. J., Medina-Garcia, G., Sotelo-Ruiz, E. D., & Ramirez-Guzman, M. E. (2022). Flooded extent and depth analysis using optical and SAR remote sensing with machine learning algorithms. *Atmosphere*, *13*(11), 1852.
86. Jenifer, A. E., & Natarajan, S. (2022). DeepFlood: A deep learning based flood detection framework using feature-level fusion of multi-sensor remote sensing images. *Journal of Universal Computer Science*, (3), 329-344.
87. Jenifer, A. E., Aparna, A., Sudha, N., & Kumar, A. (2022). AgriFloodNet: a dual patch CNN architecture for mapping flooded agricultural lands via bi-temporal multi-sensor images. *Geocarto International*, *37*(26), 13618-13637.
88. Drakonakis, G. I., Tsagkatakis, G., Fotiadou, K., & Tsakalides, P. (2022). Ombrianet—supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 2341-2356.
89. Shen, M., Zhang, F., & Wu, C. Y. (2022). Flood inundation extraction based on decision-level data fusion: A case in peru. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *10*, 133-140.
90. Jain, U., Wilson, A., & Gulshan, V. (2022). Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*.
91. Radoi, A. (2022). Multimodal satellite image time series analysis using GAN-based domain translation and matrix profile. *Remote Sensing*, *14*(15), 3734.
92. He, X., Zhang, S., Xue, B., Zhao, T., & Wu, T. (2023). Cross-modal change detection flood extraction based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, *117*, 103197.
93. Sanderson, J., Mao, H., Abdullah, M. A., Al-Nima, R. R. O., & Woo, W. L. (2023). Optimal fusion of multispectral optical and SAR images for flood inundation mapping through explainable deep learning. *Information*, *14*(12), 660.
94. Popandopulo, G., Illarionova, S., Shadrin, D., Evteeva, K., Sotiriadi, N., & Burnaev, E. (2023). Flood extent and volume estimation using remote sensing data. *Remote Sensing*, *15*(18), 4463.

95. Rawat, S., & Saini, R. (2023, March). Delineation of flooded areas using A synergistic combination of optical and SAR imagery. In *2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)* (pp. 87-92). IEEE.
96. Zhang, C., Wang, R., Chen, J. W., Li, W., Huo, C., & Niu, Y. (2023, July). 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* (pp. 5443-5446). IEEE.
97. Wang, Z., Wang, X., & Li, G. (2023, July). 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* (pp. 6454-6457). IEEE.
98. Yadav, R., Nascetti, A., & Ban, Y. (2023, July). Context-aware change detection with semi-supervised learning. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* (pp. 5754-5757). IEEE.
99. Iselborn, K., Stricker, M., Miyamoto, T., Nuske, M., & Dengel, A. (2023). On the importance of feature representation for flood mapping using classical machine learning approaches. *arXiv preprint arXiv:2303.00691*.
100. Garg, S., Feinstein, B., Timnat, S., Batchu, V., Dror, G., Rosenthal, A. G., & Gulshan, V. (2023). Cross-modal distillation for flood extent mapping. *Environmental Data Science*, *2*, e37.
101. Tavus, B., & Kocaman, S. (2023). Glcm features for learning flooded vegetation from SENTINEL-1 and SENTINEL-2 data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *48*, 601-607.
102. Albertini, C., Gioia, A., Iacobellis, V., Petropoulos, G. P., & Manfreda, S. (2024). Assessing multi-source random forest classification and robustness of predictor variables in flooded areas mapping. *Remote Sensing Applications: Society and Environment*, *35*, 101239.
103. Portalés-Julià, E., Bountos, N. I., Sdraka, M., Mateo-García, G., Papoutsis, I., & Gómez-Chova, L. (2024, July). Multimodal and multitemporal data fusion for flood extent segmentation exploiting Kurosiwo and WorldFloods Sentinel datasets. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 950-953). IEEE.
104. Sekine, T., Yamanaka, A., Eda, T., Udagawa, T., & Busto, M. (2024, July). A GAN-based SAR-optical data fusion approach for enhanced flood mapping. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 8958-8962). IEEE.
105. Wang, R., Zhang, C., Chen, C., Hao, H., Li, W., & Jiao, L. (2024). A multi-modality fusion and gated multi-filter U-Net for water area segmentation in remote sensing. *Remote Sensing*, *16*(2), 419.
106. Landuyt, L., Ivashkovych, X., & Van Achteren, T. (2024, July). Consistent Surface Water Monitoring by Fusing Sentinel-1 and-2 Through Convolutional Neural Networks. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 4701-4705). IEEE.
107. Xu, X., Xiong, M., Quan, S., Chen, J., Zhang, T., & Wei, F. (2024, July). Flood change detection based on prior feature estimation. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 10001-10005). IEEE.
108. Ahmadi, P., Valadan Zoej, M. J., Mokhtarzade, M., Kardan Halvaie, N., & Ghaderpour, E. (2025). A capsule network framework for flood mapping integrating remote sensing fusion techniques. *Environmental Research Communications*, *7*(6), 065031.

109. Portales-Julia, E., Mateo-García, G., & Gómez-Chova, L. (2025). Understanding flood detection models across Sentinel-1 and Sentinel-2 modalities and benchmark datasets. *Remote Sensing of Environment*, *328*, 114882.
110. Yailymov, B., Yailymova, H., Kolotii, A., Shelestov, A., Skakun, S., Baber, S., ... & Kussul, N. (2025). Flooded and irrigation area monitoring after the kakhovka dam disaster based on machine learning and satellite data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*, 19129-19144.
111. Krishnachaithanya, N., Sivakumar, P. B., & Deepika, T. (2025). Probabilistic flood risk mapping for Indian megacities using a scalable machine learning framework in Google Earth Engine. *Results in Engineering*, 107872.
112. Seo, M., Jung, J., & Choi, D. G. (2025). Improved flood insights: Diffusion-based SAR-to-EO image translation. *Remote Sensing*, *17*(13), 2260.
113. Carbonneau, P. E. (2025). Style transfer from sentinel-1 to sentinel-2 for fluvial scenes with multi-modal and multi-temporal image fusion. *Remote Sensing*, *17*(20), 3445.
114. Tulbure, M. G., Caineta, J., Broich, M., Gaines, M. D., Rufin, P., Thomas, L. F., ... & Hostert, P. (2025). Leveraging AI multimodal geospatial foundation models for improved near-real-time flood mapping at a global scale. *arXiv preprint arXiv:2512.02055*.
115. Negri, R. G., da Costa, F. D., da Silva Andrade Ferreira, B., Rodrigues, M. W., Bankole, A., & Casaca, W. (2025). Assessing Machine Learning Models on Temporal and Multi‐Sensor Data for Mapping Flooded Areas. *Transactions in GIS*, *29*(2), e70028.
116. Shen, Y., Yao, S., Qiang, Z., & Pei, G. (2025). SD-Mamba: A lightweight synthetic-decompression network for cross-modal flood change detection. *International Journal of Applied Earth Observation and Geoinformation*, *136*, 104409.
117. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., ... & Rolnick, D. (2025). Galileo: Learning global & local features of many remote sensing modalities. *arXiv preprint arXiv:2502.09356*.
118. Linial, O., Leifman, G., Blau, Y., Sherman, N., Gigi, Y., Sirko, W., & Beryozkin, G. (2025, February). Enhancing remote sensing representations through mixed-modality masked autoencoding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 470-479). IEEE.
119. Qin, G., Wang, S., Wang, F., Li, S., Gu, X., Chen, B., ... & Zhu, J. (2025). Rapid flood detection in global scene based on multimodal data collaboration. *International Journal of Digital Earth*, *18*(2), 2554307.
120. Savitha, M., & Meleet, M. (2025, October). Adaptive Flood Mapping with Cloud-Aware Satellite Data Selection Using AI-Based Model Adaptation. In *2025 10th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1235-1240). IEEE.
121. Zhu, J., Li, S., Zhang, Y., Liu, M., & Chen, J. (2025). Efficient Water Body Detection Based on Knowledge Distillation for SAR Imagery. *IEEE Geoscience and Remote Sensing Letters*.
122. Ibrahim, Y., Belényesi, M., Liu, C., Richter-Cserey, M., Simon, M., Szirányi, T., & Benedek, C. (2025, October). Inland Excess Water (IEW) Monitoring Using Sentinel-1/2: A SplitClass Segmentation and Temporal Gap-Filling Approach. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2875-2885). IEEE.

123. Agrawal, A., Khan, A., & Raman, B. (2025, August). Multi-View Data Fusion in Feature and Decision Spaces for Flood Inundation Mapping. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2663-2667). IEEE.
124. Lu, X., & Weng, Q. (2026). Semantic segmentation of single SAR imagery leveraging historical Sentinel-1 and Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, *146*, 105114.
125. Ahmadi, P., Valadan Zoej, M. J., Mokhtarzade, M., Kardan, N., Ahmadi, P., & Ghaderpour, E. (2026). TLE-FEDformer: A Frequency-Domain Transformer Framework for Multi-Sensor Multi-Temporal Flood Inundation Mapping. *Remote Sensing*, *18*(6), 895.
126. Talreja, J., Gebre, T. S., & Hashemi-Beni, L. (2026). FLoRA: Fusion-Latent for Optical Reconstruction and Flood Area Segmentation via Cross-Modal Multi-Task Distillation Network. *IEEE Transactions on Geoscience and Remote Sensing*.
127. Liu, Y., Wang, X., Kong, X., Ye, Y., Xu, X., Zhao, B., ... & Wen, Q. (2026). Machine learning–based flood inundation mapping using fused optical and SAR remote sensing: a case study of the Miyun flood. *Frontiers in Earth Science*, *14*, 1805240.
128. Hasi, J. M., Nakshi, S. S., & Sultana, M. (2026, April). Explainable Multi-Modal Ensemble Learning for Flood Prediction from SAR and Optical Satellite Imagery. In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)* (pp. 1-6). IEEE.
129. Luo, D., Jia, P., Jin, J., Bao, L., & Liu, W. (2026, April). Multimodal deep learning framework for remote sensing disaster recognition. In *Second International Conference on Remote Sensing Technology and Image Processing (RSTIP 2025)* (Vol. 14171, pp. 305-311). SPIE.
130. Puspitasari, D. I., Noersasongko, E., Soeleman, M. A., & Supriyanto, C. (2026). Hybrid SAR-Optical Remote Sensing for Flood Inundation Mapping: Feature Contribution Analysis in a Wetland Environment. *Statistics, Optimization & Information Computing*, *16*(1), 619-640.
131. Agrawal, A., Raman, B., & Khan, A. (2026). Divergence Guided Joint Diffusion-Convolution Network for Multimodal Flood Inundation Mapping. *IEEE Transactions on Geoscience and Remote Sensing*.
132. Hu, X., Wang, C., Zhu, D., & Li, D. (2026). Filling gaps in optical satellite images using sar imagery for flood mapping: case studies in China, Spain, and Somalia. *Geocarto International*, *41*(1), 2719974.
133. Chiriaco, G., Barco, L., Bragagnolo, A., Rossi, C., & Arnaudo, E. (2026). GEOID-Flood: A Large-Scale Multi-Modal Benchmark Dataset for Flood Segmentation. *arXiv preprint arXiv:2608.02315*.
134. Feng, H., Xiao, C., Lin, R., Chen, Y., Liang, L., & Lin, B. (2026). CISRMamba: Cross-Modal Interaction and Scan-Routing Mamba for Multi-Sensor Flood Inundation Mapping. *Sensors*, *26*(16), 5224.
135. Vinaykumar, V. N., Anusha, K., Nandeesh, G. S., Varadaraj, R., Naveen, N., & Lohith, M. S. (2026). DisasterChangeNet: Boundary-Preserving Deep Change Detection in Remote Sensing for Disaster Management. *Sensing and Imaging*, *27*(1), 149.
136. Yao, Q., Meng, Q., & Gong, Y. (2026, July). Based on a Cloud-Guided Dual-Stream MambaVision Framework for Flood Detection: A Case Study of the Miyun Reservoir Area. In *2026 33rd International Conference on Geoinformatics (Geoinformatics)* (pp. 1-7). IEEE.

137. Ahmed, S. S., Nowreen, S., & Rahman, M. S. (2026). To Remove or Not to Remove Clouds: A Comparative Analysis and Fusion of Raw SAR and Synthetic NDWI for Overcast Water Segmentation. *arXiv preprint arXiv:2608.17398*. .
138. Youssef, A., Moussa, O., Elsharkawy, A., Azouz, A., & Tarek, M. (2026). A Multi-Sensor Semi-Supervised and Unsupervised Framework for Post-Disaster Flood and Building Damage Assessment: The Case of the Derna Dam Collapse. *International Journal of Engineering and Geosciences, 11*(3), 572–596.
139. Gebre, T. S., Talreja, J., & Hashemi-Beni, L. (2026). Advanced Flood Prediction with Physics-Guided Deep Learning: Combining UNet, FNO, and SAR/Optical Imagery. *arXiv preprint arXiv:2606.06524*.
140. Naradasu, S. K., Basetty, N. K., Nakka, A., Mallela, B., Grandhe, D., & Basaiahgari, C. S. (2026, January). Deep Learning Approaches for Satellite-Based Flood Detection and Mapping. In *2026 IEEE International Conference on Emerging Computing and Intelligent Technologies (ICoECIT)* (pp. 1-6). IEEE.
141. Vaishnavi, G., Suriyasri, S., & Jagatha, S. (2026, March). Flood Vision: A Deep Learning Framework for Real Time Flood Detection From Multispectral Satellite Imagery. In *2026 8th International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1793-1798). IEEE.
142. Ebrahimi, E., Jamshidpour, N., & Homayouni, S. (2026, April). Enhancing Flood Mapping Using Integrated SAR and Optical Image Fusion with U-Net++. In *2026 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)* (pp. 31-35). IEEE.
143. Chen, Z., Xiang, D., Zhao, J., Li, J., Jiang, Y., Wu, Y., & Wen, X. (2026). Fusion of Sentinel-1 and Sentinel-2 Image Time Series for Rapid Flood Mapping based on Deep Learning Method. IAHR APD Congress.
144. Arul, A. P., Keerthika, M., Kowshikan, S., Ahmed, S., Fawaz, R. M. A., & Saravanakumar, S. (2026, March). MST-DeepLabV3+: An Integrated Deep Learning System for Real-Time Satellite-Based Flood Extent Mapping With Production-Grade Deployment. In *2026 International Conference on Future and Advanced Computing Technologies (ICFACT)* (pp. 1-10). IEEE.
145. Rijal, S., Banjade, S. S., Rai, N., Mandal, A. K., & Sharma, S. (2026). Improving flood detection in the Himalayas: Evaluating SAR-optical data fusion with machine learning in a mountainous river basin of Nepal. *Journal of the Indian Society of Remote Sensing*.
146. Kaushik, S., Maurya, L., & Tellman, B. (2026). Prithvi-Complimentary Adaptive Fusion Encoder (CAFE): Unlocking full-potential for flood inundation mapping. *arXiv*.
147. Ranjan, R., et al. (2026). SPEAR: Self-supervised sample efficient pixel-level multi-modal spectral fusion for Earth observation applications. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 1445–1454). IEEE.
148. Muszynski, M., et al. (2022). Flood event detection from Sentinel 1 and Sentinel 2 data: Does land use matter for performance of U-Net based flood segmenters? In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 4860–4867). IEEE.
149. Wagner, W., et al. (2026). The fully-automatic Sentinel-1 global flood monitoring service: Scientific challenges and future directions. *Remote Sensing of Environment, 333*, 115108.
150. Tsyganskaya, V., Martinis, S., Marzahn, P., & Ludwig, R. (2018). Detection of Temporary Flooded Vegetation Using Sentinel-1 Time Series Data. *Remote Sensing*, *10*(8), 1286.
151. Risling, A., Lindersson, S., & Brandimarte, L. (2024). A comparison of global flood models using Sentinel-1 and a change detection approach. *Natural Hazards, 120*, 11133–11152.

152. Misra, A., White, K., Nsutezo, S. F., et al. (2025). Mapping global floods with 10 years of satellite radar data. *Nature Communications, 16*, 5762.
153. Yang, Q. (2026). Operational SAR flood mapping as a full-stack systems problem: An AI-enabled perspective. *Frontiers in Water, 8*, 1871753.
154. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 62–66.
155. Uddin, K., Matin, M. A., & Meyer, F. J. (2019). Operational flood mapping using multi-temporal Sentinel-1 SAR images: A case study from Bangladesh. *Remote Sensing, 11*(13), 1581.
156. Aziz, M. A., et al. (2022). Delineating flood zones upon employing synthetic aperture data for the 2020 flood in Bangladesh. *Earth Systems and Environment, 6*(3), 733–743.
157. Singha, M., et al. (2020). Identifying floods and flood-affected paddy rice fields in Bangladesh based on Sentinel-1 imagery and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing, 166*, 278–293.
158. Bonafilia, D., Tellman, B., Anderson, T., & Kalaitzidis, C. (2020). Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 835–845). IEEE.
159. Tian, D., Wang, L., Liu, H., Cohen, S., & Shu, S. (2026). DeepSAR Flood Mapper: global flood mapping on google earth engine cloud platform using MLP deep learning model with Sentinel-1 SAR imagery and HAND topographic data. *GIScience & Remote Sensing*, *63*(1), 2612306.
160. Silwal, A., Subedi, A., Tamrakar, R., Dahal, K., Dahal, D., Ekpetere, K. O., & Zhran, M. (2026). A comprehensive review of machine learning and deep learning methods for flood inundation mapping. *Earth*, *7*(2), 44.
161. Jung, J., et al. (2026). Towards global mapping of dynamic surface water extents using Sentinel-1 SAR data. *Remote Sensing of Environment, 337*, 115326.
162. Portalés-Julià, E., et al. (2025). Understanding flood detection models across Sentinel-1 and Sentinel-2 modalities and benchmark datasets. *Remote Sensing of Environment, 328*, 114882.
163. উপজেলা রৌমারী. (2026). এক নজরে রৌমারী | পাতা. রৌমারী উপজেলা. https://rowmari.kurigram.gov.bd/pages/static-pages/69770c4cc4774958d7c007db
164. Sarker, M. H., et al. (2003). Rivers, chars and char dwellers of Bangladesh. *International Journal of River Basin Management, 1*(1), 61–80.
165. Islam, M. R., et al. (1999). The Ganges and Brahmaputra rivers in Bangladesh: Basin denudation and sedimentation. *Hydrological Processes, 13*(17), 2907–2923.
166. Bangladesh Water Development Board, Flood Forecasting and Warning Centre (FFWC), "Flood forecasting and warning services," n.d. [Online]. Available: https://www.ffwc.gov.bd/ (station-level danger-level and historical highest-water-level records cited in Section 3.2.1 were drawn from FFWC public station data; exact retrieval date and station IDs are recorded in the project data-provenance log rather than reproduced in full here).
167. Pekel, J.-F., et al. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature, 540*(7633), 418–422.
168. Olofsson, P., et al. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment, 148*, 42–57.

169. Miah, M. R. A. (2022). *Study on agricultural cropping pattern change in a water stress area of North-West region of Bangladesh: A case study in Sapahar Upazilla under Naogaon District*. BRAC University.
170. Shariot-Ullah, M. (2016). *Climate change and the Boro rice phenology in Rajshahi: Scenarios of future crop water demand for Boro rice due to climate change in Rajshahi, north-western district of Bangladesh* [Master's thesis, Wageningen University]. Wageningen University & Research.
171. Khanam, M., & Biswas, A. (2023). Determination of growth stages of some rice varieties as affected by sowing time. *Bangladesh Rice Journal, 26*(2), 41–49.
172. Islam, M. R., & Sikder, S. (2011). Phenology and degree days of rice cultivars under organic culture. *Bangladesh Journal of Botany, 40*(2), 149–153.
173. Madhu, U., et al. (2018). Influence of sowing date on the growth and yield performance of wheat (Triticum aestivum L.) varieties. *Archives of Agriculture and Environmental Science, 3*(1), 89–94.
174. Sikder, S. (2009). Accumulated heat unit and phenology of wheat cultivars as influenced by late sowing heat stress condition. *Journal of Agriculture & Rural Development*, 59–64.
175. Zhao, J., Chini, M., Pelich, R., Matgen, P., Hostache, R., Cao, S., & Wagner, W. (2020). Deriving exclusion maps from C-band SAR time-series: An additional information layer for SAR-based flood extent mapping. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-1–2020* (pp. 395–400). Copernicus Publications.