

Event Aware Flood Mapping for Agricultural Landscapes: A Robustness Oriented Comparison of Deep Learning and Machine Learning in the Arkansas 2025 Flood Event

Manh-Dung Vu

vdung8417@gmail.com

National Taiwan University <https://orcid.org/0009-0001-6585-2007>

Gia-Hien Tran

Hanoi – Amsterdam Highschool for the Gifted <https://orcid.org/0009-0008-8764-5632>

Ming-Che Hu

mchu@ntu.edu.tw

National Taiwan University <https://orcid.org/0000-0002-5288-2936>

Research Article

Keywords: flood mapping, agricultural exposure, event aware predictors, U Net, Random Forest, uncertainty analysis, Arkansas floodplain

Posted Date: April 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9384304/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Event Aware Flood Mapping for Agricultural Landscapes: A Robustness Oriented Comparison of Deep Learning and Machine Learning in the Arkansas 2025 Flood Event

Manh-Dung Vu^{1*}, Gia-Hien Tran², Ming Che Hu^{1*}

¹ Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan

² Hanoi-Amsterdam High School for the Gifted, No. 1 Hoang Minh Giam Street, Cau Giay District, Hanoi 100000, Vietnam

*Corresponding author. E mail: mchu@ntu.edu.tw

Abstract

Accurate flood mapping is essential for rapid impact assessment in agricultural landscapes, where inundation can interrupt production, damage rural infrastructure, and complicate recovery. Recent research has shown that deep learning can support flood delineation from remotely sensed data, but the literature still reports recurring weaknesses in robustness assessment, uncertainty interpretation, and transparent benchmarking against simpler machine learning baselines. In this study, we developed an event aware flood mapping framework for the Arkansas 2025 flood event by integrating terrain, antecedent water conditions, rainfall dynamics, and hydrologic proximity within a spatial learning workflow. We evaluated progressively richer predictor configurations using a U Net based convolutional neural network and assessed robustness through repeated seed experiments under fixed train, validation, and test splits. A Random Forest benchmark was introduced to provide a transparent classical machine learning reference. Mean probability, uncertainty, and agreement maps were also generated to characterize spatial confidence patterns. The results showed that adding rainfall substantially improved performance relative to the more complex advanced deep learning configuration, increasing mean best validation IoU by 7.3%, mean test IoU by 15.8%, and mean test F1 by 12.4%, while also reducing the variability of test IoU, test F1, and threshold selection. The rainfall based U Net achieved a best validation IoU of 0.278 ± 0.015 , a test IoU of 0.264 ± 0.014 , and a test F1 of 0.417 ± 0.018 . The more complex configuration that added distance to river and Focal Tversky optimization did not deliver a consistent gain, with a best validation IoU of 0.259 ± 0.011 , a test IoU of 0.228 ± 0.033 , and a test F1 of 0.371 ± 0.044 . The Random Forest benchmark was unexpectedly strong, reaching a validation IoU of 0.608 at the selected threshold and a test IoU of 0.351 with a test F1 of 0.519, equivalent to gains of 33.0% in test IoU and 24.5% in test F1 relative to the best deep learning configuration. Feature importance analysis identified elevation, slope, rainfall peak, and event total rainfall as dominant contributors. Spatial uncertainty was concentrated along transitional flood margins, while agreement maps highlighted coherent high confidence cores. Cropland overlay analysis further demonstrated that the framework can support first pass agricultural exposure assessment. Rather than promoting model complexity for its own sake, this study shows that event aware predictors, robustness oriented evaluation, and explicit benchmark comparison are central to credible geospatial flood intelligence in agricultural environments.

Keywords: *flood mapping; agricultural exposure; event aware predictors; U Net; Random Forest; uncertainty analysis; Arkansas floodplain*

1. Introduction

Floods remain among the most damaging natural hazards worldwide. Their effects now extend far beyond short periods of emergency disruption. In many vulnerable regions, repeated inundation changes land use over time, erodes livelihood security, and shapes the pathways through which recovery unfolds [1,2]. Agricultural landscapes are especially exposed to these pressures. In such settings, a flood map is not merely a record of water presence. It is the first spatial layer from which crop exposure, access constraints, and recovery priorities are inferred [7,8].

Earth observation has transformed flood monitoring by making large event footprints visible across broad areas that are often difficult to reach in the field. Synthetic aperture radar has been especially valuable because it remains effective in darkness and under cloud cover, two conditions that commonly coincide with the most critical phases of a flood event [2,3,4]. The most dependable flood products, however, rarely emerge from imagery alone. Their quality depends on whether terrain controls, antecedent conditions, and event scale forcing are incorporated in a physically sensible manner [3,4,6].

Deep learning has become increasingly prominent in flood mapping because convolutional architectures can learn nonlinear spatial relations from stacked raster predictors and generate dense outputs at pixel scale with high efficiency once training is complete [5,9,10]. That progress has been real, but it has not resolved several practical weaknesses in the literature. Recent reviews point to uneven benchmarking practice, only limited attention to robustness across repeated runs, and weak communication of spatial confidence [3,5]. Even in studies that report strong flood detection performance, differences in training data, split design, and uncertainty treatment often make it difficult to judge whether a model is genuinely dependable or simply favoured by one particular experimental configuration [12,13,14].

These concerns become more serious when flood delineation is expected to support agricultural interpretation downstream. Recent studies linking inundation products to cropland damage or crop condition demonstrate clear operational value, yet they also show that downstream inference is only as credible as the upstream flood layer [7,8]. A study that compares model structures, predictor sets, and benchmark baselines carefully is scientifically more useful than one that reports only a single best score. The practical issue is not limited to where flooding is predicted. It also concerns whether that prediction remains stable enough to be trusted when the output is later intersected with cropland and other exposure layers.

The present study addresses this gap through a robustness oriented event aware framework for the Arkansas 2025 flood event. The framework was designed to keep the event itself at the centre of predictor construction, rather than relying only on static susceptibility. Deep learning was evaluated against a transparent classical baseline instead of being presumed superior from the outset. Performance was interpreted through repeated seed analysis and spatial uncertainty products, not through one favourable run alone. This framing aligns with recent calls for more rigorous operational flood intelligence that balances physical plausibility, predictive power, and interpretability [5,33,35].

The study pursued a set of closely connected objectives. It examined the added value of event aware rainfall predictors relative to a static baseline, while also testing whether additional complexity, represented here by distance to river and Focal Tversky optimization, could deliver gains that remained stable across repeated training. The selected deep learning configurations were then compared against a Random Forest benchmark [11], and the resulting inundation maps were further assessed for their ability to support cropland exposure screening. The study therefore advances a more balanced perspective on flood mapping

research, one in which robustness, benchmark honesty, and downstream usefulness are treated as core scientific criteria rather than optional add ons.

2. Materials and Methods

2.1 Study context and analytical design

The Arkansas 2025 flood event provided the empirical setting for this analysis. A single analytical domain was maintained throughout predictor preparation, model training, evaluation, and final mapping so that all stages rested on the same spatial frame. The study did not aim to reproduce flood hydraulics through direct simulation. Its purpose was to learn the spatial correspondence between event aware geospatial predictors and a reference inundation map, and then examine whether that correspondence remained reliable under repeated training. This approach suits rapid flood mapping well, especially when multiple raster sources must be brought into one modelling framework and the resulting predictions are intended for subsequent exposure screening [5,11].

The workflow was organised as a continuous sequence rather than a set of isolated steps. It began with the construction of the predictor stack, which combined terrain information, long term surface water context, pre event water conditions, and rainfall descriptors. That input base was then converted into patch based datasets for training, validation, and testing. Model comparison followed, covering a static baseline U Net, a rainfall enriched U Net, a more complex advanced U Net, and a Random Forest benchmark. The final part of the analysis focused on interpretation. Threshold sweeps were used to define decision boundaries, repeated seed experiments were used to assess robustness, and the resulting outputs were examined through spatial error maps, uncertainty maps, agreement maps, and cropland exposure overlay. This workflow kept the analysis close to operational flood mapping practice, while retaining a structure that could be reproduced and evaluated systematically.

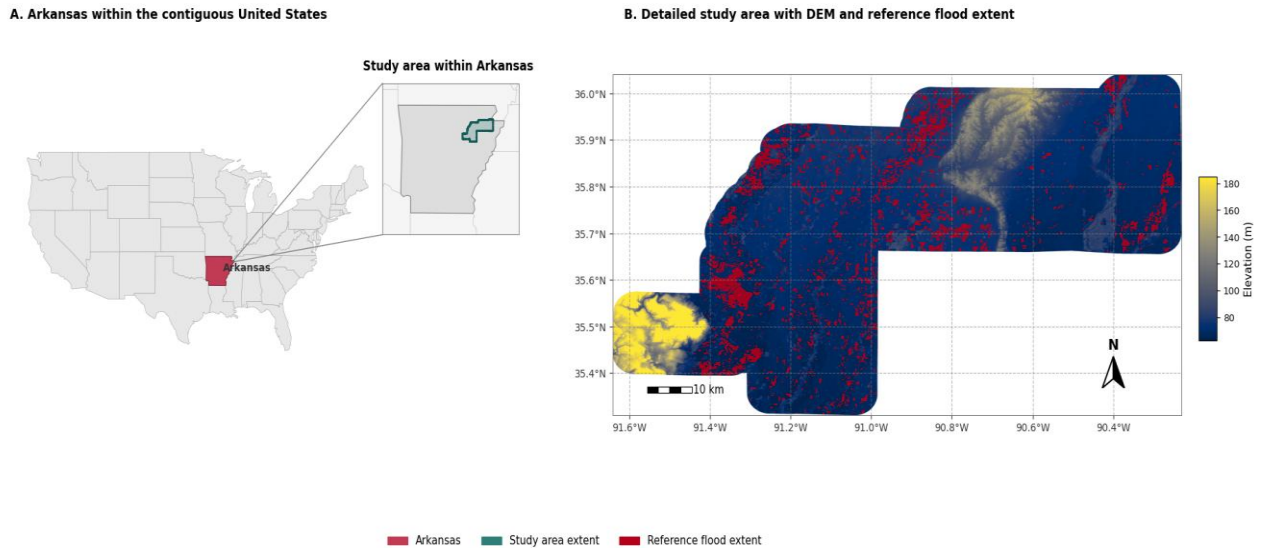


Figure 1. Study area and case setting of the Arkansas 2025 flood event. Panel A locates Arkansas within the contiguous United States and indicates the position of the study area within the state. Panel B provides a closer view of the study area. The digital elevation model is used as the topographic base, while the reference flood extent is shown in red. Geographic coordinates are provided to aid spatial orientation. A segmented scale bar and a north arrow are also included for map reading.

2.2 Input geospatial predictors

Topography formed the physical background of the predictor stack. Elevation was taken from the Copernicus GLO 30 digital elevation model, and slope was derived from the same source. This dataset provides globally available 30 m elevation information and is widely used in broad scale terrain analysis and environmental screening [17]. To represent longer term hydrologic context, the analysis used the Joint Research Center Global Surface Water product. Because this dataset was derived from millions of Landsat scenes, it offers a well established basis for identifying historical water occurrence and persistence [18]. Cropland extent was drawn from ESA WorldCover 2021 at 10 m resolution. WorldCover provides Sentinel based global land cover information, and its class definitions have been independently evaluated against other recent high resolution global products [19,20,42].

Event forcing was characterized with precipitation metrics derived from NASA GPM IMERG Version 07 [15,16]. IMERG is widely used in hydrologic and disaster studies because it provides spatially consistent multisatellite precipitation estimates with sub daily temporal coverage. The rainfall variables were chosen to capture both antecedent wetness and direct event forcing, covering moisture accumulation before the flood together with total rainfall and peak one day intensity during the event itself. This structure allowed the models to see more than a single rainfall snapshot and instead respond to a short temporal sequence of moisture build up and triggering conditions.

Two additional layers were used to describe water state before and beyond the event. Permanent water identified areas of routinely persistent water that should not be interpreted as event flooding [18]. Pre event water captured water already present before the flood and helped distinguish newly inundated areas from antecedent wetness. The advanced configuration also included a distance to river proxy derived from permanent water. This variable was not intended to reproduce hydraulic routing in a strict sense. Its role was simpler and more practical: to provide an indicator of hydrologic proximity that remains conceptually relevant in floodplain analysis [3,26,27,28].

Table 1. *Input geospatial predictors, data sources, spatial role, and citation base.*

Predictor	Purpose in the model	Primary source
Elevation and slope	Terrain control on storage, connectivity, and lateral spread	Copernicus DEM [17]
Permanent water	Masking and long term hydrologic context	JRC Global Surface Water [18]
Pre event water	Antecedent water state before the event	Event specific preprocessing
Rainfall seven day total	Antecedent wetness signal	GPM IMERG V07 [15,16]
Rainfall three day total	Short term preconditioning signal	GPM IMERG V07 [15,16]
Event rainfall total	Integrated event forcing	GPM IMERG V07 [15,16]

Peak one day rainfall	Short duration high intensity forcing	GPM IMERG V07 [15,16]
Distance to river	Hydrologic proximity proxy for the advanced model	Derived in this study
Cropland mask	Exposure overlay for agricultural screening	ESA WorldCover [19,20]

2.3 Reference flood map and supervised learning target

The supervised learning target was a reference flood map prepared for the Arkansas 2025 event and aligned to the common analytical grid. Its role in this study was to provide a stable binary segmentation target rather than a full physical description of depth, velocity, or flood duration. Flood mapping studies often combine sources that carry different forms and levels of uncertainty, and the quality of the target product can therefore place a practical ceiling on achievable segmentation performance [5,33]. The reference layer is therefore interpreted here as an operational delineation target, useful for mapping evaluation and exposure screening, but not as complete hydraulic truth.

All predictor rasters were resampled and normalized to a common grid before patch extraction. Robust percentile based normalization was applied to continuous layers, whereas binary masks were kept unchanged. This choice reduced the influence of extreme values without erasing the spatial contrasts that mattered for flood mapping.

2.4 Model configurations

Three supervised deep learning configurations were evaluated. The modeling strategy developed step by step from a static baseline built with elevation, slope, permanent water, and pre event water. Rainfall metrics were then added to this predictor set, giving rise to the rainfall model. A further extension introduced distance to river and used Focal Tversky loss in training so that class imbalance could be treated more aggressively [29,30]. Each configuration added one conceptual component to the previous one, which allowed the contribution of each predictor family to be interpreted through direct comparison.

A Random Forest classifier was also implemented as the benchmark baseline [11]. For structured geospatial predictors, Random Forest remains a strong classical reference, particularly in settings where nonlinear relationships matter and predictor scales are not fully uniform. Its feature importance outputs also make interpretation more transparent. Including this benchmark was necessary for the logic of the paper. If deep learning truly adds value, that value should remain visible when tested against a strong classical comparator rather than against a weak baseline.

Table 2. Model configurations evaluated in the study.

Configuration	Predictor set	Learning method	Purpose
Static baseline	Elevation, slope, permanent water, pre event water	U Net with BCE Dice loss	Baseline susceptibility plus antecedent water
Rainfall model	Static predictors plus four rainfall metrics	U Net with BCE Dice loss	Event aware mapping configuration

Advanced model	Rainfall model plus distance to river	U Net with Focal Tversky loss	Test whether added complexity improves robustness
Random Forest benchmark	Rainfall predictor set	Random Forest classifier	Transparent machine learning comparator

2.5 U Net architecture and training strategy

The deep learning models were built around a U Net architecture [9], itself a canonical extension of fully convolutional segmentation principles [10]. For geospatial segmentation, U Net remains a practical choice because it captures broader contextual information in the contracting path and then restores spatial detail in the expanding path. Skip connections further help retain local structure that might otherwise be lost during encoding. This architectural balance is one reason U Net and related variants have been used so widely in flood detection research, including studies based on benchmark SAR datasets and near real time applications [5,21,22,23,24,25].

Training was performed on patch based samples extracted from the harmonized predictor stack. The same train, validation, and test partition was kept fixed across repeated seed experiments, which made it possible to examine the effect of random initialization without conflating it with changes in data splitting. Binary cross entropy combined with Dice style overlap optimization was used for the static baseline and rainfall model, drawing on a family of overlap aware segmentation objectives that are widely used for class imbalanced tasks [31,32]. The advanced model used Focal Tversky loss to place more emphasis on difficult minority pixels [29,30]. In this study, however, the loss function was treated as an empirical hypothesis rather than as an assumed improvement.

2.6 Threshold sweep, repeated seed robustness, and uncertainty analysis

Model outputs were treated as probability maps. For each trained model, a validation threshold sweep was used to identify the threshold that maximized validation IoU, and that selected threshold was then transferred to the test data for final binary evaluation. This procedure was preferred over arbitrary thresholding because it keeps the decision rule explicit and reproducible.

Repeated seed analysis was then conducted under fixed train, validation, and test splits to characterize sensitivity to random initialization and data order. Performance across repeated runs was summarized with measures of central tendency and dispersion, covering validation behaviour, test behaviour, and the selected decision threshold. Spatial confidence was examined in map space by identifying where predicted flood probability remained stable, where variability accumulated, and where repeated runs converged or diverged in their delineation of inundation. This strategy does not yield Bayesian posterior uncertainty in the strict sense, but it does provide an empirical view of predictive stability, which is useful in light of the broader literature on calibration, ensemble uncertainty, and uncertainty propagation in hazard modelling [12,13,14,33,34,35].

2.7 Agricultural exposure overlay

To move beyond pure delineation accuracy, predicted flood extent was intersected with the cropland mask derived from ESA WorldCover [19,20]. The objective was not to estimate crop loss directly. Instead, the goal was to test whether the mapping workflow could support a first pass agricultural exposure

screening layer. This application step follows the logic of recent work in which flood delineation is used as the spatial basis for crop damage screening and post event agricultural assessment [7,8].

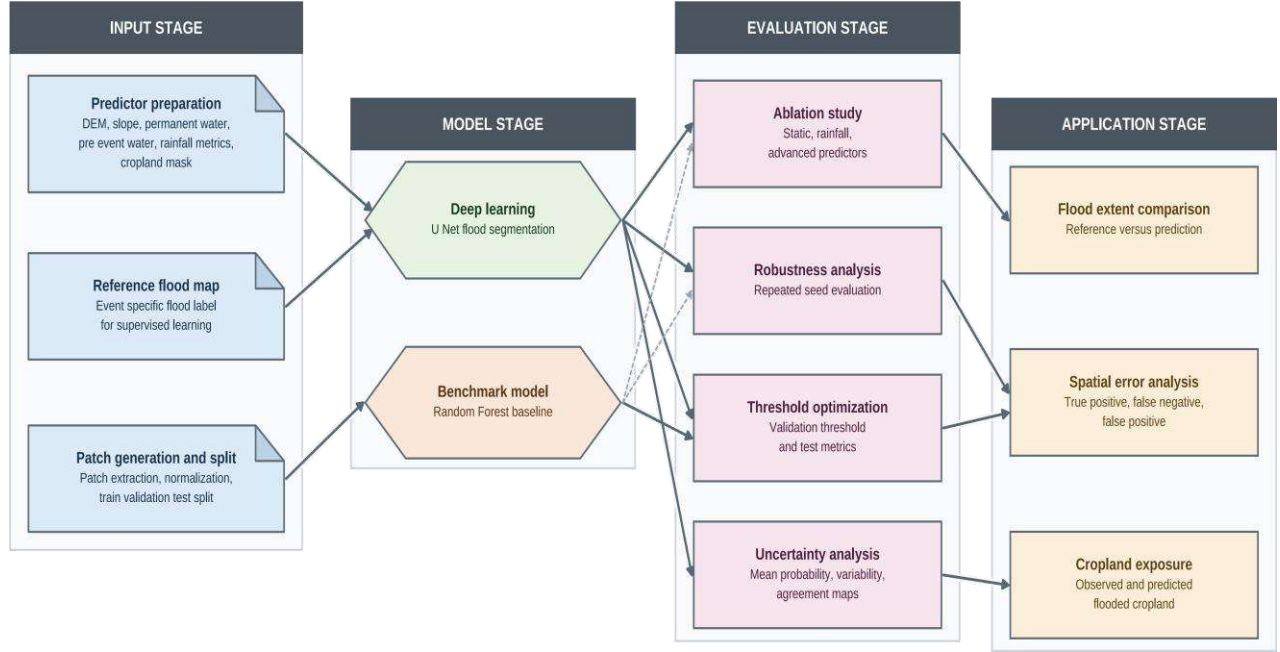


Figure 2. Conceptual workflow of the study, from predictor preparation and patch generation to deep learning, benchmark comparison, uncertainty analysis, and cropland exposure assessment.

3. Results

3.1 Ablation study of the deep learning configurations

Table 3 and Figure 3 together show a clear hierarchy among the deep learning configurations. The static baseline performed worst, indicating that terrain and antecedent water alone were insufficient to reconstruct the observed inundation pattern of the Arkansas 2025 event. Once rainfall metrics were introduced, performance improved materially. This outcome is physically credible. Static predictors describe where flooding is plausible, but event rainfall helps explain why flooding actually occurred in a particular episode and how strongly the signal should express itself across the floodplain.

As reported in Table 3, the rainfall model increased mean best validation IoU from 0.259 to 0.278, mean test IoU from 0.228 to 0.264, and mean test F1 from 0.371 to 0.417. Figure 3 shows the same pattern in visual form, with the rainfall model occupying a higher and tighter performance range across repeated seeds. In relative terms, those gains correspond to 7.3%, 15.8%, and 12.4%, respectively. This is precisely the kind of difference that matters in applied flood mapping, because the better model is not only more accurate on average but also easier to trust across repeated training runs.

Threshold behaviour adds a second layer of evidence. The rainfall model selected a mean threshold of 0.300 ± 0.030 , whereas the advanced model required 0.433 ± 0.115 . That means the advanced configuration was about 3.8 times more variable in threshold selection. The validation to test drop also favoured the rainfall model. Its mean best validation IoU decreased from 0.278 to 0.264 on the test set, an absolute gap of 0.014 or roughly 5.0% relative to validation performance. The advanced model dropped from 0.259 to 0.228, an absolute gap of 0.031 or about 12.0%. Table 3 and Figure 3 therefore point to the

same conclusion: the more complex configuration not only scored lower on average, but also generalized less cleanly from validation to test data.

Table 3. Main performance summary for the repeated seed deep learning experiments under fixed train, validation, and test splits.

Model	Best validation IoU	Test IoU	Test F1	Best threshold
Rainfall model, mean $\pm$ SD	0.278 ± 0.015	0.264 ± 0.014	0.417 ± 0.018	0.300 ± 0.030
Advanced model, mean $\pm$ SD	0.259 ± 0.011	0.228 ± 0.033	0.371 ± 0.044	0.433 ± 0.115

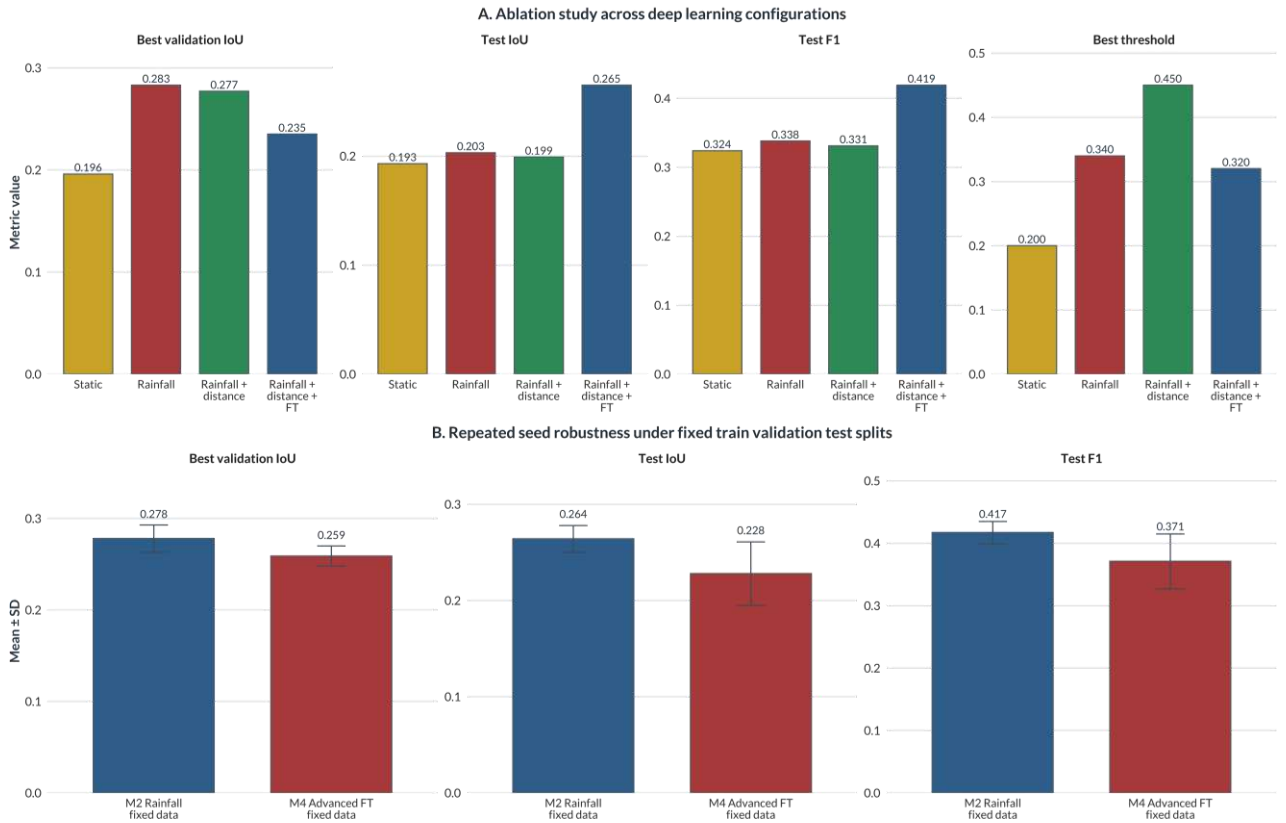


Figure 3. Comparative performance across the ablation study and repeated seed robustness analysis for the deep learning configurations.

3.2 Random Forest benchmark

Table 4 shows that the Random Forest benchmark did more than remain competitive, as it delivered the strongest performance on the held out test set. Using the same event aware rainfall predictor set as the rainfall based U Net, Random Forest reached a validation IoU of 0.608 at the selected threshold of 0.35, together with a test IoU of 0.351 and a test F1 of 0.519. Relative to the rainfall based U Net, these results correspond to gains of 0.087 in test IoU and 0.102 in test F1, or 33.0% and 24.5%, respectively, indicating that predictor engineering itself carried a large share of the usable signal and providing a demanding reference against which any added value from architectural complexity had to be demonstrated.

Table 4 also calls for a more careful reading of the benchmark result. The gap between validation IoU and test IoU for Random Forest was 0.257, equivalent to 42.3% of its validation score, which means that validation performance alone would have conveyed a more favourable picture than the held out test data support. Random Forest nevertheless remained the strongest performer on unseen data, but the size of this decline makes it clear that the benchmark should be interpreted with the same discipline applied to Table 3, with particular attention to how much performance is retained after transfer beyond the validation set.

Figure 4 sharpens the physical interpretation of that benchmark result. Elevation and slope ranked first and second, followed by rainfall peak, event total rainfall, and seven day antecedent rainfall. Pre event water and three day antecedent rainfall still contributed, whereas permanent water had the weakest influence in the selected model. The ranking matters because it shows that the classifier did not behave like a rainfall only detector. It relied first on the floodplain template defined by topography, then used rainfall metrics to identify where that template was activated during the event under study. That hierarchy gives the Random Forest result a much more defensible hydrologic reading.

Table 4. *Random Forest benchmark performance and selected threshold for the rainfall predictor set.*

Metric	Value
Best validation threshold	0.35
Validation IoU at selected threshold	0.608
Validation F1 at selected threshold	0.756
Test IoU	0.351
Test F1	0.519
Test precision	0.454
Test recall	0.607
Test accuracy	0.873

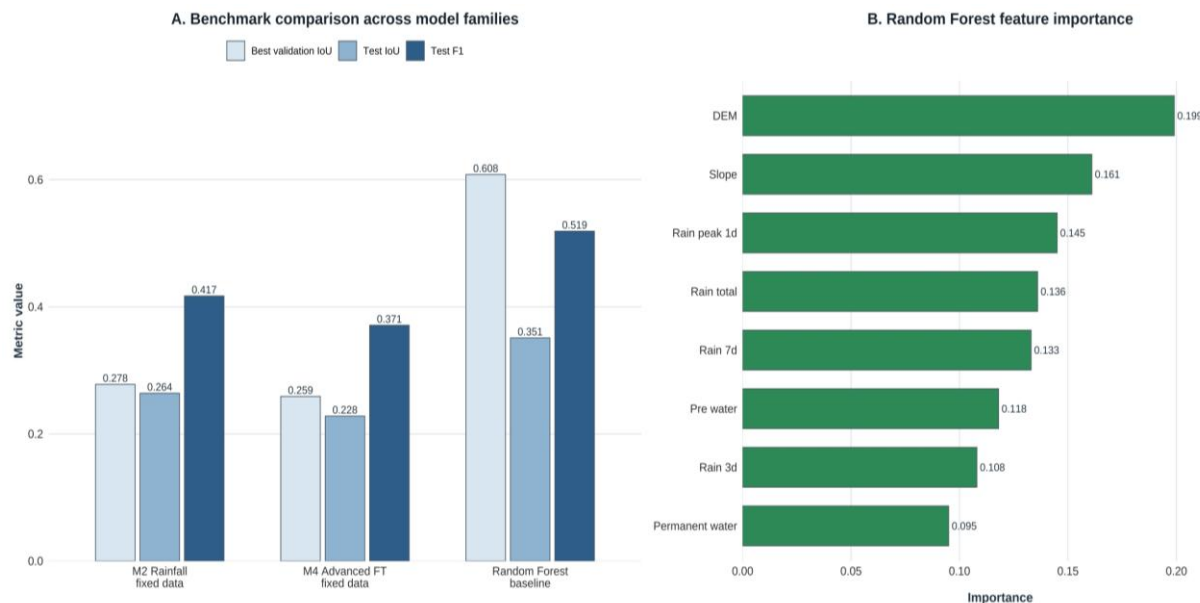


Figure 4. Random Forest feature importance for the rainfall predictor set.

3.3 Spatial flood comparison and error structure

Figures 5 and 6 help clarify why the rainfall model remained the preferred deep learning configuration. In Figure 5, both models recover the broad organization of the floodplain. The difference lies in the continuity of the mapped inundation. The rainfall based prediction preserves the main flooded body more cleanly, whereas the advanced configuration breaks more often along the margins and within secondary patches. This spatial contrast is consistent with Table 3, where the rainfall model retained an advantage of 0.036 in test IoU and 0.046 in test F1. The map that appears more coherent is therefore also the stronger performer numerically.

In Figure 6, the error structure reveals a further distinction between the two models. For the rainfall model, error pixels cluster more heavily near flood boundaries, which suggests that much of the remaining disagreement is concentrated in transition zones. In the advanced model, false negatives and false positives are distributed more diffusely across the floodplain interior. That pattern is harder to justify scientifically, because it points to weaker internal coherence rather than uncertainty confined to the edge of the inundated zone. This distinction matters in review, since some ambiguity at flood margins is understandable in event based mapping, whereas broad interior fragmentation is much more difficult to defend.

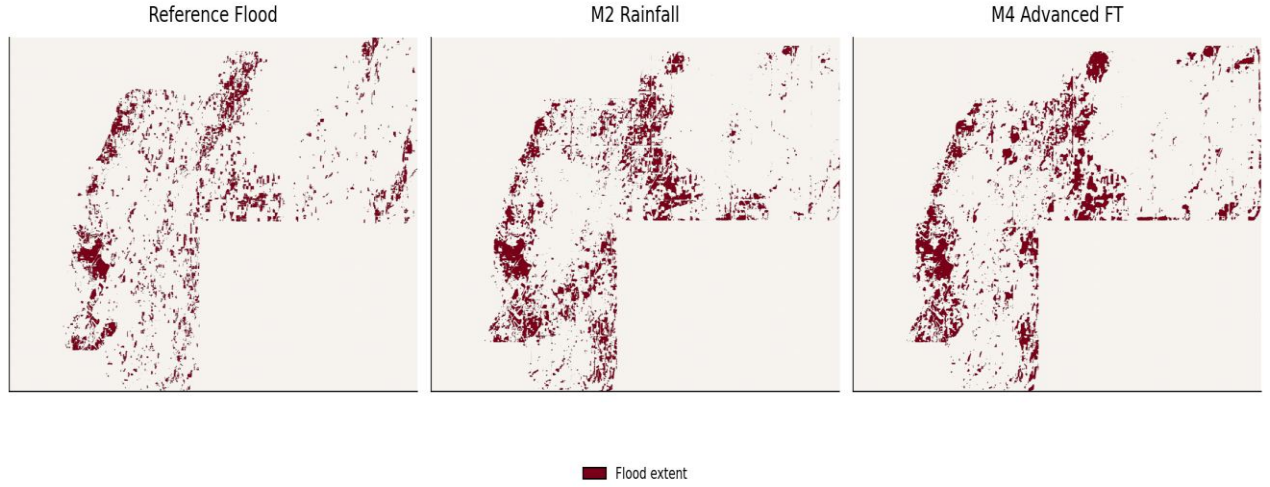


Figure 5. Spatial comparison between the reference flood extent and the model derived inundation maps for the Arkansas 2025 flood event. The figure presents the reference flood map alongside the rainfall based prediction and the advanced model output.

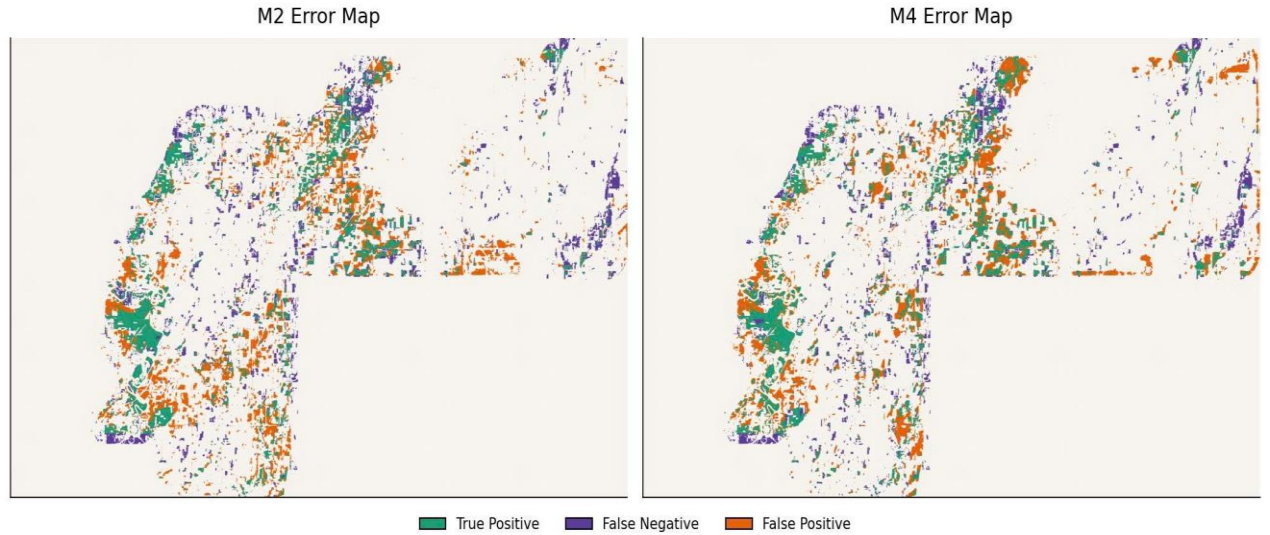


Figure 6. Spatial distribution of prediction errors in the rainfall based model and the advanced configuration, evaluated against the reference flood map. Green denotes true positives. Purple marks false negatives. Orange marks false positives.

3.4 Uncertainty and agreement maps

Figure 7 provides information that cannot be recovered from scalar metrics alone. The mean probability panel indicates a stable central flood signal, whereas the uncertainty panel shows that disagreement is concentrated mainly along flood margins and in fragmented peripheral areas. This spatial pattern clarifies how the model behaves across the floodplain. The flooded core is reproduced with considerable consistency, while ambiguity becomes more visible near class transitions and in locations where inundation is broken into smaller patches.

The agreement panel becomes especially informative when considered together with Table 3. For the rainfall model, the standard deviation of the selected threshold was only 0.030, and the corresponding standard deviation of test IoU was 0.014. Figure 7 shows how those relatively small dispersions appear in map space: repeated runs converge over the principal inundated zones, whereas divergence is concentrated mainly where the flood boundary is less distinct. The uncertainty pattern is therefore structured in a physically plausible way, rather than being scattered randomly across the mapped floodplain.

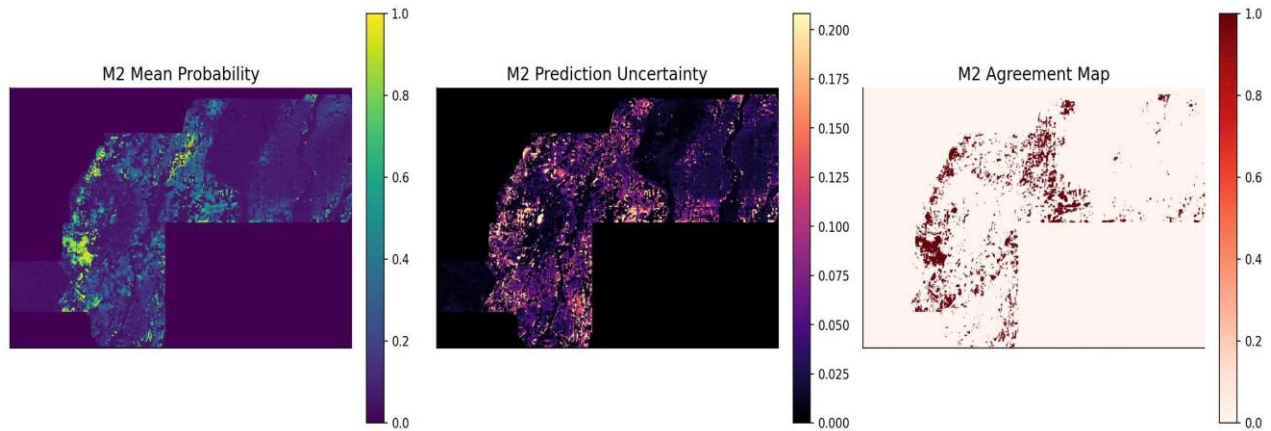


Figure 7. Spatial patterns of uncertainty and agreement in the rainfall based model across repeated seed experiments. The figure presents the mean predicted flood probability, the spatial distribution of prediction uncertainty, and the level of agreement among runs. Together, these panels distinguish areas of stable inundation mapping from parts of the Arkansas floodplain where model ambiguity remains higher.

3.5 Cropland exposure assessment

Figure 8 shows why the segmentation results matter beyond model comparison alone. The rainfall model was carried forward to cropland overlay because it emerged as the strongest and most stable deep learning configuration. After intersection with cropland, the predicted flooded pattern still preserved the main spatial structure visible in the observed flooded cropland layer, most clearly in the larger contiguous agricultural blocks adjacent to the principal inundation belt. Although this result should not be interpreted as evidence of exact crop loss, it indicates that a model with a test IoU of 0.264 and a test F1 of 0.417 can still provide a spatially useful first screen of agricultural exposure.

This point is especially relevant for applied remote sensing and hazard research. Figure 8 is useful not because it claims parcel scale agronomic precision, but because it shows that the level of error reported in Table 3 remains low enough for the dominant geography of exposure to survive the overlay step. Operational value, in this case, depends less on perfect delineation than on whether the main clusters of exposed cropland remain visible with sufficient stability, and Figure 8 suggests that this condition was met.

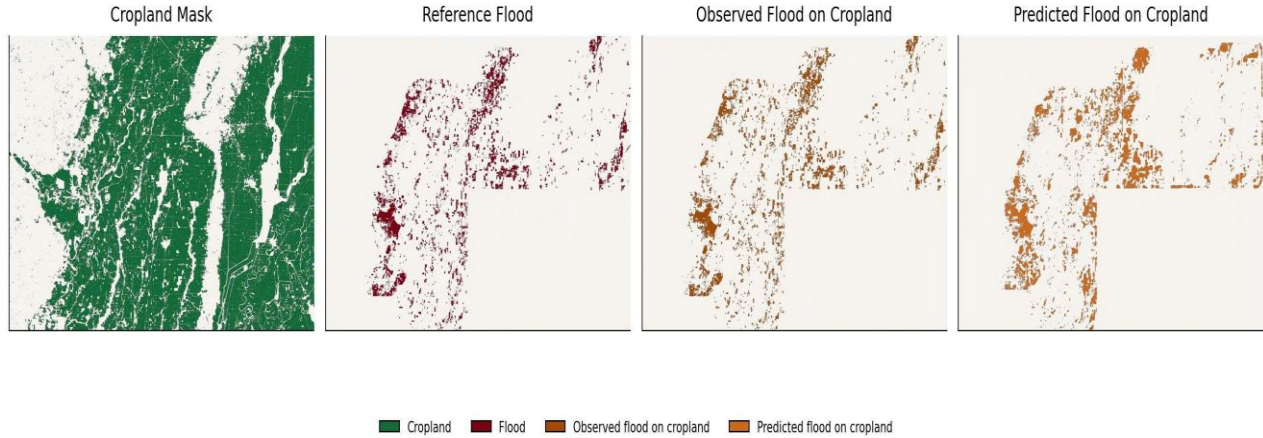


Figure 8. Agricultural exposure to flooding in the Arkansas 2025 event. The panels show the cropland mask, the reference flood extent, observed flooded cropland derived from the overlap between cropland and the reference flood map, and predicted flooded cropland derived from the overlap between cropland and the best performing model output.

4. Discussion

4.1 Why rainfall mattered most

The strongest and most consistent result of the study was the gain obtained by adding event rainfall metrics. Terrain and long term water layers describe predisposition, but they do not encode the timing or magnitude of the forcing that activates inundation in a specific event. This pattern can be seen in Table 3 and in Figures 3, 5, and 7. Once rainfall was added, mean test IoU rose to 0.264 and mean test F1 rose to 0.417, while the spatial outputs became more continuous and the cross seed uncertainty remained concentrated mainly at flood margins. Recent reviews of SAR flood mapping and deep learning flood mapping repeatedly emphasize that auxiliary variables are most useful when they capture physically meaningful event context rather than being added only for feature volume [3,4,5]. The result also aligns with recent efforts to enhance satellite based inundation products with topographic and process aware information so that flood delineation becomes more event specific and more usable for downstream interpretation [6,39,40,41].

4.2 Why more complexity did not guarantee better performance

The advanced model was built on reasonable ideas. Distance to river should, in principle, inject additional hydrologic structure, and Focal Tversky loss should offer a sharper response to class imbalance [29,30]. Yet the repeated seed results showed that these additions did not convert into a robust gain. This is precisely why methodological discipline matters. More complexity is useful only when it captures additional signal more effectively than it amplifies sensitivity in training or threshold selection. In the present case, the rainfall model achieved the better balance between representational richness and empirical stability, and the evidence is not subtle. Its test IoU was higher by 0.036, its test F1 was higher by 0.046, and its threshold variability was much lower.

This conclusion is strengthened by the distinction between validation behaviour and test behaviour. The advanced model was not uniformly erratic, and its best validation IoU variability was even slightly smaller. But the model was clearly less stable where it mattered most, namely in test IoU, test F1, and threshold

selection. The threshold standard deviation of 0.115 for the advanced configuration versus 0.030 for the rainfall model is especially revealing, because it suggests that the more complex setup required a much less stable decision boundary. The outcome therefore supports a practical lesson that extends beyond this case study. Loss functions and auxiliary predictors should be justified by reproducible gains, not by their conceptual sophistication alone [5,12,13,14,32].

4.3 What the Random Forest result means

The strength of the Random Forest benchmark is one of the most informative outcomes of the manuscript. Table 4 shows that event aware geospatial predictors can carry substantial predictive power even before a deep architecture exploits local spatial context. With a test IoU of 0.351 and a test F1 of 0.519, the benchmark outperformed the rainfall based U Net by a meaningful margin. Figure 4 helps explain why this result is scientifically plausible, as the feature ranking follows a physically coherent pattern in which topography and event rainfall jointly dominate the prediction process. High performance, therefore, does not automatically validate greater model complexity, since it may instead indicate that feature design is already carrying much of the scientific burden.

The benchmark nevertheless requires careful interpretation. Random Forest achieved the highest held out test metrics, but it also displayed a notable decline from validation to test performance, showing both that classical machine learning remains highly competitive in structured geospatial settings [11,39] and that robustness diagnostics remain necessary regardless of model family. In future work, repeated seed or repeated resampling analysis should also be extended to the benchmark, so that comparisons across model families rest on matched stability diagnostics rather than on headline scores alone.

4.4 Uncertainty as a scientific output rather than an afterthought

One recurring criticism of geospatial artificial intelligence in hazard applications is that it often delivers visually persuasive outputs without an equally careful account of confidence. This concern is well documented in both machine learning and flood modeling literature [12,13,14,33,35]. The present study therefore treated uncertainty products as part of the main analytical output rather than as an optional supplement. Figure 7 is central here, because it reveals where the model is stably confident and where that confidence weakens.

Although the uncertainty maps presented here are not Bayesian posterior estimates, repeated seed uncertainty and agreement products still remain informative because they show how prediction stability is distributed under repeated training. In practice, this information can guide how strongly an analyst should trust a local delineation when the flood layer is later combined with exposure data. Their value lies in the spatial pattern itself. The main inundated core continues to show strong agreement across runs, whereas less stable behavior is concentrated near the margins and in fragmented parts of the floodplain. Under these conditions, uncertainty is more than a visual supplement, since it materially changes how the mapped result should be interpreted.

4.5 Implications for agricultural landscapes

Agricultural exposure assessment was included because it provides a concrete test of usefulness beyond segmentation accuracy. Recent crop oriented flood studies have shown that remote sensing based inundation products can support timely post event assessment, but they also make clear that the quality of the flood layer remains the key upstream constraint [7,8]. Figure 8 is therefore important for more than presentation. It shows that the selected deep learning output preserved the main geography of flooded

cropland strongly enough to support a first pass exposure screen, even though the flood model itself was far from perfect in a strict segmentation sense.

In agricultural environments, flood intelligence is most useful when developed as a progressive analytical process that begins with stable inundation mapping and then moves into exposure overlay before any attempt is made to estimate actual damage from crop calendars, field boundaries, flood duration, or water depth. That broader damage oriented step lies beyond the scope of the present study, but stating the sequence clearly helps keep the analysis scientifically precise without losing its operational relevance. Figure 8 also takes on a more explicit methodological role in this context, because the overlay serves as the first operational check of whether the mapped flood signal is informative enough to support agricultural decision making.

4.6 Limitations and future work

The present study was conducted within a tightly defined event setting, which made it possible to examine the contribution of event rainfall and classical benchmarking under controlled conditions. Some limitations nevertheless remain important for how the framework should be interpreted and extended. The rainfall inputs were coarser than several of the static predictors, and this mismatch in spatial resolution may have reduced boundary sharpness in more heterogeneous parts of the floodplain. The hydrologic proximity variable was also intentionally simple. It should be understood as a practical indicator of relation to the local drainage structure, not as a physical representation of flood propagation.

Several next steps follow directly from these constraints. Broader validation against curated reference products and standardized datasets is needed to support transparent comparison across modelling approaches [21,22,23,24,25]. A richer representation of landscape controls would also be valuable, especially through variables such as HAND or MERIT Hydro that can better reflect hydrogeomorphological structure and compound inundation processes [36,37,38]. Closer integration with hybrid operational frameworks is another logical direction, particularly where rapid statistical prediction needs to be interpreted alongside process based modelling and explicit treatment of uncertainty [40,41,43]. Developments along these lines would provide a stronger test of how well the current framework performs beyond the present event and under a wider range of flood conditions.

5. Conclusions

This study developed and evaluated a predictive framework for the Arkansas 2025 flood event with a particular focus on robustness. The clearest result was the role of rainfall. Once event rainfall information was introduced, the deep learning models performed more reliably across repeated random initializations. By contrast, the more complex configuration, which added distance to river and Focal Tversky optimization, did not deliver gains that remained stable from one run to the next.

The Random Forest result also changed how the model comparison should be interpreted. Its strong performance showed that well chosen geospatial predictors already captured a large share of the inundation signal. This point matters because it shifts the emphasis of the study away from whether a deep model can produce a plausible flood map and toward whether it continues to add value when tested against a strong classical benchmark rather than a weak baseline.

The study also shows that flood mapping should not be judged from a single favourable run. Stability across repeated initialization, spatial uncertainty patterns, and the cropland overlay all contribute to a more credible evaluation. For agricultural landscapes, a useful inundation framework needs more than acceptable

accuracy, since it must remain stable under repeated training, be interpretable in map space, and preserve enough spatial structure to support downstream use. The main contribution of this work lies not in arguing that greater complexity will always improve performance, but in showing how event aware predictors, transparent benchmarking, and uncertainty informed interpretation can be combined within a more reliable workflow for geospatial flood analysis.

Data Availability Statement

The datasets used for model development were assembled from publicly accessible geospatial products and event specific processing steps. Public source products include NASA GPM IMERG, Copernicus DEM, JRC Global Surface Water, and ESA WorldCover [15,16,17,18,19]. Event specific processed rasters, training patches, model outputs, and figure source files should be archived by the authors and deposited in a public repository at the time of submission or revision.

Conflict of Interest Statement

The authors declare no competing interests.

References

1. Tellman B, Sullivan JA, Kuhn C, Kettner AJ, Doyle CS, Brakenridge GR, et al. Satellite imaging reveals increased proportion of population exposed to floods. *Nature*. 2021;596:80–86. Available from: <https://doi.org/10.1038/s41586-021-03695-w>
2. Misra A, White K, Nsutezo SF, Straka W III, Lavista J, et al. Mapping global floods with 10 years of satellite radar data. *Nature Communications*. 2025;16:5762. Available from: <https://doi.org/10.1038/s41467-025-60973-1>
3. Shen X, Wang D, Mao K, Anagnostou EN, Hong Y. Inundation Extent Mapping by Synthetic Aperture Radar: A Review. *Remote Sensing*. 2019;11(7):879. Available from: <https://doi.org/10.3390/rs11070879>
4. Amitrano D, Di Martino G, Di Simone A, Imperatore P. Flood Detection with SAR: A Review of Techniques and Datasets. *Remote Sensing*. 2024;16(4):656. Available from: <https://doi.org/10.3390/rs16040656>
5. Bentivoglio R, Isufi E, Jonkman SN, Taormina R. Deep learning methods for flood mapping: a review of existing applications and future research directions. *Hydrology and Earth System Sciences*. 2022;26:4345–4378. Available from: <https://doi.org/10.5194/hess-26-4345-2022>
6. Betterle A, Salamon P. Water depth estimate and flood extent enhancement for satellite based inundation maps. *Natural Hazards and Earth System Sciences*. 2024;24:2817–2836. Available from: <https://doi.org/10.5194/nhess-24-2817-2024>
7. Wen C, Sun Z, Li H, Han Y, Gunasekera D, Chen Y, et al. Flood Mapping and Assessment of Crop Damage Based on Multi Source Remote Sensing: A Case Study of the 7.27 Rainstorm in Hebei Province, China. *Remote Sensing*. 2025;17(5):904. Available from: <https://doi.org/10.3390/rs17050904>
8. Rahman MS, Di L, Yu E, Lin L, Yu Z. Remote Sensing Based Rapid Assessment of Flood Crop Damage Using Novel Disaster Vegetation Damage Index. *International Journal of Disaster Risk Science*. 2021;12:674–689. Available from: <https://doi.org/10.1007/s13753-020-00305-7>
9. Ronneberger O, Fischer P, Brox T. U Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer Assisted Intervention, MICCAI 2015*. Cham: Springer; 2015. p. 234–241. Available from: https://doi.org/10.1007/978-3-319-24574-4_28

10. Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 3431 3440. Available from: <https://doi.org/10.1109/CVPR.2015.7298965>
11. Breiman L. Random Forests. Machine Learning. 2001;45(1):5 32. Available from: <https://doi.org/10.1023/A:1010933404324>
12. Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Advances in Neural Information Processing Systems. 2017;30. Available from: <https://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles>
13. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017;70:1321 1330. Available from: <https://proceedings.mlr.press/v70/guo17a.html>
14. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems. 2019;32. Available from: <https://doi.org/10.48550/arXiv.1906.02530>
15. NASA Global Precipitation Measurement Mission. IMERG V07 Technical Documentation. 2023. Available from: <https://gpm.nasa.gov/resources/documents/imerg-v07-technical-documentation>
16. Google Earth Engine Data Catalog. NASA GPM L3 IMERG V07. Available from: https://developers.google.com/earth-engine/datasets/catalog/NASA_GPM_L3_IMERG_V07
17. Google Earth Engine Data Catalog. Copernicus DEM GLO 30: Global 30 m Digital Elevation Model. Available from: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_DEM_GLO30
18. Pekel JF, Cottam A, Gorelick N, Belward AS. High resolution mapping of global surface water and its long term changes. Nature. 2016;540:418 422. Available from: <https://doi.org/10.1038/nature20584>
19. Zanaga D, Van De Kerchove R, Daems D, De Keersmaecker W, Brockmann C, Kirches G, et al. ESA WorldCover 10 m 2021 v200. 2022. Available from: <https://doi.org/10.5281/zenodo.7254221>
20. Xu P, Tsendbazar NE, Herold M, de Bruin S, Koopmans M, Birch T, et al. Comparative validation of recent 10 m resolution global land cover maps. Remote Sensing of Environment. 2024;311:114316. Available from: <https://doi.org/10.1016/j.rse.2024.114316>
21. Bonafilia D, Tellman B, Anderson T, Issenberg E. 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. 2020. p. 835 845. Available from: <https://doi.org/10.1109/CVPRW50498.2020.00113>
22. Bai Y, Wu W, Yang Z, Yu J, Zhao B, Liu X, et al. 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. 2021;13(11):2220. Available from: <https://doi.org/10.3390/rs13112220>
23. Katiyar V, Tamkuan N, Nagai M. Near Real Time Flood Mapping Using Off the Shelf Models with SAR Imagery and Deep Learning. Remote Sensing. 2021;13(12):2334. Available from: <https://doi.org/10.3390/rs13122334>
24. Ghosh B, Garg S, Motagh M, Martinis S. Automatic Flood Detection from Sentinel 1 Data Using a Nested UNet Model and a NASA Benchmark Dataset. PFG Journal of Photogrammetry, Remote Sensing and Geoinformation Science. 2024;92:1 18. Available from: <https://doi.org/10.1007/s41064-024-00275-1>
25. Wu X, Zhang Z, Xiong S, Zhang W, Tang J, Li Z, et al. A Near Real Time Flood Detection Method Based on Deep Learning and SAR Images. Remote Sensing. 2023;15(8):2046. Available from: <https://doi.org/10.3390/rs15082046>

26. Twele A, Cao W, Plank S, Martinis S. Sentinel 1 based flood mapping: a fully automated processing chain. *International Journal of Remote Sensing*. 2016;37(13):2990 3004. Available from: <https://doi.org/10.1080/01431161.2016.1192304>
27. Huang X, Wang C, Li Z. A near real time flood mapping approach by integrating social media and post event satellite imagery. *Annals of GIS*. 2018;24(2):113 123. Available from: <https://doi.org/10.1080/19475683.2018.1450787>
28. Cian F, Marconcini M, Ceccato P. Normalized Difference Flood Index for rapid flood mapping: Taking advantage of EO big data. *Remote Sensing of Environment*. 2018;209:712 730. Available from: <https://doi.org/10.1016/j.rse.2018.03.006>
29. Salehi SSM, Erdogmus D, Gholipour A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: Wang Q, Shi Y, Suk HI, Suzuki K, editors. *Machine Learning in Medical Imaging*. Cham: Springer; 2017. p. 379 387. Available from: https://doi.org/10.1007/978-3-319-67389-9_44
30. Abraham N, Khan NM. A Novel Focal Tversky loss function with improved Attention U Net for lesion segmentation. *arXiv*. 2018. Available from: <https://doi.org/10.48550/arXiv.1810.07842>
31. Milletari F, Navab N, Ahmadi SA. V Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision. 2016. p. 565 571. Available from: <https://doi.org/10.48550/arXiv.1606.04797>
32. Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising Dice and cross entropy based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*. 2022;95:102026. Available from: <https://doi.org/10.1016/j.compmedimag.2021.102026>
33. Muñoz DF, de Moel H, Dottori F, Muis S, Haigh ID, Winsemius HC, et al. Quantifying cascading uncertainty in compound flood modeling with linked process based and machine learning models. *Hydrology and Earth System Sciences*. 2024;28:2531 2555. Available from: <https://doi.org/10.5194/hess-28-2531-2024>
34. Dhara S, Dang T, Parial K, Lu XX. Accounting for Uncertainty and Reconstruction of Flooding Patterns Based on Multi Satellite Imagery and Support Vector Machine Technique: A Case Study of Can Tho City, Vietnam. *Water*. 2020;12(6):1543. Available from: <https://doi.org/10.3390/w12061543>
35. de Moel H, van Alphen J, Aerts JCJH. Effect of uncertainty in land use, damage models and inundation depth on flood damage estimates. *Natural Hazards*. 2011;58(1):407 425. Available from: <https://doi.org/10.1007/s11069-010-9675-6>
36. Eilander D, Couasnon A, Leijnse T, Ikeuchi H, Winsemius HC. A globally applicable framework for compound flood hazard modeling. *Natural Hazards and Earth System Sciences*. 2023;23:823 857. Available from: <https://doi.org/10.5194/nhess-23-823-2023>
37. Yamazaki D, Ikeshima D, Sosa J, Bates PD, Allen GH, Pavelsky TM. MERIT Hydro: A High Resolution Global Hydrography Map Based on Latest Topography Datasets. *Water Resources Research*. 2019;55:5053 5073. Available from: <https://doi.org/10.1029/2019WR024873>
38. Nobre AD, Cuartas LA, Momo MR, Pinheiro A, Tomasella J, Rodriguez DA, et al. Height Above the Nearest Drainage: a hydrologically relevant new terrain model. *Journal of Hydrology*. 2011;404(1 2):13 29. Available from: <https://doi.org/10.1016/j.jhydrol.2011.03.051>
39. Amer R. Machine Learning Driven Rapid Flood Mapping for Tropical Storm Imelda Using Sentinel 1 SAR Imagery. *Remote Sensing*. 2025;17(11):1869. Available from: <https://doi.org/10.3390/rs17111869>
40. Bao D, Xue ZG, Hiatt M, Xu K, Harris CK, Trepanier JC, et al. A machine learning based prediction to map framework for rapid and accurate spatial flood prediction. *npj Natural Hazards*. 2025;2:71. Available from: <https://doi.org/10.1038/s44304-025-00122-2>

41. Wagner W, Martinis S, Schlaffer S, Matgen P, Hostache R, Chini M, et al. The fully automatic Sentinel 1 Global Flood Monitoring service: scientific challenges and future directions. *Remote Sensing of Environment*. 2026;310:115108. Available from: <https://doi.org/10.1016/j.rse.2025.115108>
42. Venter ZS, Barton DN, Chakraborty T, Simensen T, Singh G. Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. *Remote Sensing*. 2022;14(16):4101. Available from: <https://doi.org/10.3390/rs14164101>
43. Nevo S, Morin E, Rosenthal AG, Metzger A, Barshai C, Weitzner D, et al. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*. 2022;26:4013 4033. Available from: <https://doi.org/10.5194/hess-26-4013-2022>