UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SHORT-RANGE FORECASTING OF FLASH FLOOD WARNINGS WITH DEEP LEARNING

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By EVAN CHLADNY

Norman, Oklahoma

2026

SHORT-RANGE FORECASTING OF FLASH FLOOD WARNINGS WITH DEEP LEARNING

A THESIS APPROVED FOR THE

SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Aaron Hill, Chair

Dr. Nathan Snook

Dr. Kathy Pegion

Table of Contents

Abstract

Flash flood forecasting remains a major operational challenge because impacts develop rapidly and are influenced by highly localized rainfall and hydrologic processes. This thesis investigates whether machine learning can complement existing forecast guidance by providing short-term, storm-scale probabilistic predictions of National Weather Service (NWS) flash flood warning (FFW) issuance. A U-Net convolutional neural network was developed to predict the likelihood of FFW issuance at 0-, 1-, and 2-hour lead times over the contiguous United States.

Data included Multi-Radar Multi-Sensor (MRMS) precipitation, Flooded Locations and Simulated Hydrographs (FLASH) Coupled Routing and Excess Storage (CREST) maximum unit streamflow, High Resolution Rapid Refresh (HRRR) meteorological fields, land cover, and optional HRRR quantitative precipitation forecasts (QPF) from the warm seasons in 2021-2025.

Verification results demonstrated strong regional variability in model performance, with higher skill in areas of frequent flash flood occurrence and reduced skill in lower-frequency regions. Spatial tolerance analysis showed that the models often correctly identified general regions of flooding but struggled to reproduce the precise spatial extent of FFW polygons, often exceeding the polygon bounds. Finally, improved metrics with alternative NWS flood products suggest that the models may be capturing a broader flood signal instead of strictly flash floods.

Overall, these results indicate that machine learning can provide useful storm-scale probabilistic guidance for flash flooding. However, this prediction problem requires careful consideration of the data being used, as errors are strongly influenced by the limitations of FFW polygons as a training target. This suggests that future work should explore more physically representative labels, such as creating a flood/flash flood practically perfect dataset.

Chapter 1: Introduction and Background

1.1 Motivation

Flash flooding is a meteorological and hydrological phenomenon characterized by rapid accumulation of surface water that poses a direct threat to life and property (National Weather Service 2026). These events are usually caused by intense rainfall over a small area for a duration of up to six hours (Doswell et al. 1996; American Meteorological Society 2026; National Weather Service 2026). Once critical rainfall thresholds are reached, the effects of flash flooding can begin in less than an hour to almost instantaneously (Archer and Fowler 2018).

Due to the highly-localized and rapid nature of these events, flash flood prediction remains a critical challenge in operational meteorology (Soares et al. 2025). An example of the challenges faced when forecasting flash floods occurred on July 4th, 2025. During this event, the Hill Country region of central Texas experienced historic and catastrophic flash flooding that resulted in the loss of over one hundred lives, with a majority of fatalities in Kerr County (NOAA National Centers for Environmental Information 2026). Intense rainfall associated with a slow-moving convective system rapidly overwhelmed the region's river basins, producing sudden and extreme rises in water levels along the Guadalupe River and its tributaries. The speed at which the flooding developed left little time for action, trapping individuals in their homes and requiring extensive evacuation efforts (Fingerhut 2025).

Forecast guidance in the days leading up to the event indicated an elevated risk for flash flooding across central Texas, including the issuance of a flood watch at 18:18 UTC July 3rd that highlighted the potential for localized rainfall totals of 5-7 inches (National Weather Service 2025). While this guidance acknowledged the potential for intense, localized rainfall, flood

watches are not intended to provide the degree of specificity to predict flooding for specific locations or river valleys. Furthermore, the Weather Prediction Center's (WPC) Extreme Rainfall Outlook (ERO) on the day before and the day of the event depicted only a slight risk of any given location exceeding flash flood guidance (Weather Prediction Center 2026), reflecting uncertainty in the evolution of the system. As the event unfolded, mesoscale storm dynamics and convective organization controlled the location and intensity of rainfall, limiting the practical lead time with which extreme impacts could be anticipated. These storm-scale dynamics occurred on time scales of minutes to tens of minutes and over spatial scales of a few tens of kilometers. Small spatial errors in numerical weather prediction when combined with the complex terrain and highly responsive watersheds of the Hill Country further complicated forecasting flash flood conditions. Thus, even when forecasters have high confidence in a flash flood event due to favorable synoptic conditions for extreme rainfall, prediction of the precise location of flash flooding remains constrained by storm-scale processes and local geography.

Beyond individual events, flash flooding remains a devastating hazard in the United States. From 1995 to 2024 flash flooding accounted for, on average, over 100 fatalities annually (National Weather Service 2024). The economic impacts are similarly substantial, with flash flooding causing $3.14 billion in property damage in 2024 alone (NOAA National Centers for Environmental Information 2026). Extreme events can periodically produce losses that exceed one billion dollars. From 2020 to 2024, flooding and flash flooding was responsible for eight separate billion-dollar disasters in the United States (NOAA National Centers for Environmental Information 2024). Collectively, these statistics demonstrate that flooding is an ongoing and costly hazard, with flash flooding in particular posing a disproportionate threat due to its rapid onset and localized nature (Terti et al. 2017).

Motivated by these predictability limitations and costly impacts, this thesis explores the potential for machine learning (ML) to complement existing forecast guidance by enhancing short-term flash flood prediction.

1.2 Current Tools

Central to flash flood monitoring in operations is the radar-based Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation (QPE) system (Zhang et al. 2016). MRMS provides near-real-time estimations of rainfall intensity and accumulation, with updates every two minutes, allowing forecasters to continuously assess where heavy precipitation is occurring. This high-resolution observational product is currently used in National Weather Service operations and forms the foundation of precipitation situational awareness (Gerard et al. 2021).

Observed rainfall fields from MRMS are often interpreted in conjunction with Flash Flood Guidance (FFG), which quantifies the amount of rain over a specific time and area required to fill small streams to full capacity (Gerard et al. 2021). Further precipitation past FFG is therefore associated with flash flooding. Unlike MRMS, which updates at sub-hourly timescales, FFG is updated up to four times a day at the NOAA River Forecast Center's discretion (Clark et al. 2014). While FFG remains a widely used reference tool in operations, there are limitations that can reduce its accuracy in diagnosing flash flood potential, such as not considering land use (Seo et al. 2013).

To provide a more physically representative depiction of ongoing hydrologic response, more complex systems have been developed to supplement purely rainfall-based guidance. One such system is Flooded Locations and Simulated Hydrographs (FLASH), which employs distributed

hydrologic modeling forced by real-time MRMS QPE (Yussouf et al. 2020). Unlike FFG, which is updated four times per day and remains static between updates, FLASH is dynamically updated as new MRMS observations become available. This allows for changes in soil moisture and antecedent rainfall to be immediately represented in each update. FLASH has been implemented operationally, with established critical thresholds to help forecasters know when flash flood warnings are likely needed (Gerard et al. 2021). Importantly, FLASH is driven exclusively by QPE and does not incorporate quantitative precipitation forecasts (QPF). As a result, it does not provide probabilistic guidance or meaningfully extend flash flood forecast lead time (Yussouf et al. 2020).

Recognizing the need to extend FLASH beyond observations, Yussouf et al. (2020) coupled the Warn-on-Forecast System (WoFS) (Stensrud et al. 2009, 2013) with FLASH to generate probabilistic flash flood guidance at storm scale. This approach demonstrated that high-resolution physics-based ensemble forecasts can provide a bridge between model output and flash flood decision support. However, the WoFS-FLASH framework remains dependent on computationally expensive modeling currently deployed in limited operational contexts.

While high-resolution numerical weather prediction (NWP) models provide valuable guidance for heavy rainfall and flash flood potential, they also have important limitations when used in isolation for flash flood forecasting. Convection-allowing models such as the High-Resolution Rapid Refresh (HRRR) and ensemble systems like WoFS can produce QPF and probabilistic guidance through ensemble spread. However, NWP models primarily provide meteorological information rather than direct estimates of flash flood potential. Forecasters must interpret model output in conjunction with observations and hydrologic guidance to assess flash

flood risk. These intermediate steps between model output and warning decisions limits the ability of NWP alone to provide actionable, storm-scale flash flood decision support.

These limitations motivate the development of approaches that combine the strengths of NWP with observational and hydrologic information in a more direct way. Machine learning offers one such pathway by learning statistical relationships among diverse sets of data to make predictions.

1.3 Machine Learning Forecasting

A range of methods have been applied to predict flooding and flash flooding with machine learning. Several papers have relied on supervised learning techniques to relate modeled rainfall and other hydrometeorological variables to flooding (Mosavi et al. 2018; Hill and Schumacher 2021; Schumacher et al. 2021; Nevo et al. 2022; Hill et al. 2024). Hill and Schumacher (2021) developed a random forest-based system to emulate EROs using predictors derived from the National Severe Storms Laboratory Weather Research and Forecasting (NSSL-WRF) model. While these forecasts targeted average recurrence interval (ARI) exceedance rather than flash flood occurrence directly, it serves as a proxy for daily flash flooding potential. Mosavi et al. (2018) uses multiple linear regression and long-short term memory (LSTM) to provide real time flood warnings to agencies and the public, particularly on large, gauged rivers. Nevo et al. (2022) explores in depth the many other papers on this topic.

Martinaitis et al. (2023) presents a shorter-term approach to flash flood prediction. In this framework, traditional numerical weather prediction, WoFS, and hydrologic modeling, FLASH, are used to simulate rainfall and streamflow. Supervised machine learning is then applied as a

postprocessing step to generate probabilistic guidance. The system produces high-resolution forecasts of flash flood risk on timescales of minutes to hours, with probabilities that communicate risk and intensity information. This method emphasizes physically based modeling, with machine learning used to translate model output into probabilistic flood guidance.

Recently, deep learning has emerged as a powerful machine learning approach in meteorology. In particular, convolutional neural networks (CNNs) are well suited for meteorological applications because they automatically learn spatial features from gridded data that are relevant to a target phenomenon, without manual identification of which patterns or predictors are most important (Chase et al. 2023). Traditional CNNs are often used for image classification tasks that produce a single output value. In contrast, a subset of CNNs known as U Network (U-Net) architectures are designed for spatial prediction problems. U-Nets preserve spatial structure through an image-to-image framework, taking in images and producing outputs of similar shape (Ronneberger et al. 2015). As a result, U-Nets excel at predicting two- and three-dimensional targets (Chase et al. 2023). Compared to traditional CNNs, U-Nets incorporate resolution changing in their learning processes, which encourages the model to learn "what" and "where" certain regions are at different points in the training process. The resolution changes can also aid in generalizing information (lower spatial scales), while also paying attention to details (higher spatial scales).

Li et al. (2024) investigated the use of U-Net and enhanced U-Net architectures (Nested-Unet, U-Net 3Plus, and SmaAt-Unet) for flood nowcasting by coupling deep learning-based precipitation nowcasting with a hydrological model. In contrast to Martinaitis et al. (2023), where machine learning is applied after the physical modeling, Li et al. (2024) uses machine

learning to generate precipitation inputs that drive a hydrologic model. In their framework, various types of U-Nets were trained to extrapolate future radar reflectivity fields at lead times of 0.5, 1, and 2 hours using sequences of Doppler radar observations. The predicted reflectivity fields were converted to rainfall estimates using a dynamic reflectivity-rainfall (Z-R) relationship. The rainfall estimates were subsequently used as input to the HEC-HMS hydrological model to simulate flood potential. Results showed that U-Net-based models produced reasonable reflectivity fields at short lead times, particularly at 0.5-1 hour. Forecast skill degraded rapidly at longer lead times due to growing uncertainty in U-Net precipitation inference, leading to substantial underestimation of peak water discharge at 2 hours. Therefore, U-Net-based models improved radar precipitation nowcasting relative to baseline methods, but rapidly increasing uncertainty with lead time ultimately limited longer-term forecast skill.

Beyond flash flood prediction, machine learning has also been used to directly predict or identify hazardous weather in real-time storm-scale systems. Within WoFS, machine learning has been used to generate probabilistic forecasts for hazards such as hail, damaging winds, and tornadoes (Flora et al. 2021; Clark and Loken 2022). In addition to WoFS, the Tornado Probability (TORP) algorithm uses machine learning for the rapid identification of tornadoes from radar data. These efforts demonstrate that machine learning can be successfully deployed in real time, storm-scale environments, providing guidance that compliments traditional numerical weather prediction or observational analysis.

Overall, prior studies demonstrate that machine learning can be applied to hazardous weather prediction in several ways. These include postprocessing physically based model output to estimate flood likelihood, using deep learning to generate precipitation inputs for hydrologic modeling, and directly predicting hazardous weather at the storm scale. Each approach reflects a

different role for machine learning within the forecasting system and is tied to a specific prediction target. In Hill and Schumacher (2021), the prediction target had to be carefully defined and optimized to align with operational standards. Li et al. (2024) showed that target selection in the temporal dimension has a significant impact on model accuracy. Flora et al. (2025) also explored this theme with ML prediction of hail from WoFS, further confirming that data choice for ML models can have big implications on their effectiveness.

1.4 Operational Motivation

Real-time FFW decisions require forecasters to interpret rapidly changing, highly localized information under substantial uncertainty. In operational settings, this is not a problem of data availability, but rather the need to quickly interpret and synthesize a large volume of observations and model guidance as events unfold. While the machine learning studies in Chapter 1.3 have demonstrated skill in predicting hazard-related outcomes, fewer efforts have focused on developing tools that directly support storm-scale flash flood warning decisions in real-time. As a result, an operational gap remains for machine learning approaches that directly target real-time flash flood prediction using information consistent with observed flash flood impacts and warning decisions.

These considerations motivate further exploration of machine learning approaches that are explicitly formulated around real-time flash flood forecasting needs. Approaches that directly target flash flood occurrence as identified operationally may offer a complementary pathway to existing precipitation- and proxy-based guidance. This thesis examines such an approach, with an emphasis on short-term, real-time, storm-scale flash flood prediction.

Chapter 2: Data

This chapter describes the datasets used to develop and evaluate a machine learning (ML) system designed to predict the likelihood of National Weather Service (NWS) flash flood warning (FFW) issuance with 0–2-hour lead time. To approximate the information available in an operational forecasting environment, the ML system integrates multiple sources of observational and numerical modeling data, including radar-derived precipitation and hydrological fields, environmental predictors, and static land cover information. Each dataset is described in detail in the following sections, along with the processing steps used to align them spatially and temporally.

2.1 Warnings

The objective of this thesis is to develop an ML framework that estimates the probability of FFW issuance and updates every 10 minutes. Rather than using a rainfall threshold or hydrologic exceedance as the learning target, historical FFWs are used because they represent operational forecaster decisions made in response to observed or anticipated flash flooding. Accordingly, the presence or absence of an active FFW is treated as the target variable for model training and evaluation. To construct a gridded FFW dataset, official NWS warning polygons were obtained from the Iowa Environmental Mesonet (IEM) for the warm season months of May through September from 2021 to 2025. The dataset was limited to polygon-based warnings within the Contiguous United States (CONUS) and included information on the time of issuance and expiration for each warning. Warning polygons were mapped onto the HRRR 3-km grid at 10-minute temporal resolution. For each 10-minute interval, the beginning of which is

considered the decision time, a given pixel was considered warned if an active warning included that pixel at any time during the 10-minute window. Active warning polygons were spatially merged into a single composite geometry using GeoPandas to limit redundant spatial operations. Grid points contained within one or more active warnings were assigned a value of 1, while all other grid points were assigned a value of 0, resulting in a binary FFW mask.

To ensure a smooth label dataset free of grid-scale artifacts resulting from polygon geometry and mapping onto the HRRR grid, spatial smoothing was applied to the binary fields. A 3x3 neighborhood majority filter was used, such that a grid cell was classified as warned if at least 5 of the 9 surrounding grid cells were also classified as warned regardless of if the point was warned previously. This smoothing step reduced single-pixel artifacts while maintaining the general spatial extent of the warnings.

2.2 MRMS

Two core components of the MRMS system were used for training the U-Net. The first predictor was MRMS precipitation rate, and the second was FLASH Coupled Routing and Excess Storage (CREST) Unit Streamflow. Precipitation rate is a radar estimated product that is produced every two minutes (Zhang et al. 2016). Unit Streamflow then uses this precipitation information to estimate the maximum discharge of surface water normalized by upstream basin area every ten minutes (Gerard et al. 2021). As a general guideline, operational flash flood forecasters consider unit streamflow values exceeding 2 $m^3\ s^{-1}\ km^{-2}$ to be an indicator of potential flash flooding (Gerard et al. 2021), making it directly relevant to FFW issuance.

MRMS precipitation rate and CREST unit streamflow data between 00 UTC 1 May to 23 UTC 30 September, 2021-2025, were downloaded as hourly archive files from the IEM database. These archive files were converted to compressed NetCDF files using a lossy compression algorithm which retained precision appropriate for each variable of interest. MRMS data were then interpolated from their native 1km grid (Gerard et al. 2021) to the coarser (3 km horizontal grid spacing) HRRR grid used by the ML model. Following interpolation, a precomputed CONUS mask was expanded outward by 45 km beyond the MRMS data boundary, effectively creating a buffer zone near the edges of the dataset. Convolutional neural networks can produce artifacts near data boundaries since they rely on surrounding pixels (kernel) to make a prediction about a single pixel (Chase et al. 2023). By expanding the mask in this way, any edge-related artifacts are confined to regions outside the retained analysis domain, which is the true CONUS.

Once reformatted to match the warning dataset, a post-processing pass was performed on the 2-minute MRMS precipitation rate dataset to address CONUS-wide missing data at specific timesteps. Short gaps ($2 \leq \Delta t \leq 10$ minutes) were filled using the time-weighted average of the preceding and subsequent precipitation rate fields. Longer gaps ($\Delta t > 10$) were assigned NaN values (marking them as missing data). Rainfall accumulations were then calculated from the 2-minute precipitation rate data.

To create accumulated precipitation totals, each 2-minute precipitation rate was converted into a 2-minute increment of precipitation accumulation by taking the product of the precipitation rate and time step length in hours:

$$Increment\ (mm) = Precipitation\ Rate\ \left(\frac{mm}{hr}\right) * 2\ (minute) * \frac{1}{60}\ \left(\frac{hr}{minute}\right)$$

If a time step contained missing data for a grid cell, the accumulation for that grid cell was also flagged as missing data (assigned a value of NaN). Accumulations were computed for multiple rolling windows spanning 0-1, 1-3, 3-6, 6-12, 12-24 hours prior to the decision time. These interval breakpoints were selected to align with flash flooding timescales. The bins were defined as non-overlapping in order to preserve temporal information about when the rainfall occurred. Accumulated outputs were written to files in 10-mintue increments to match the Unit Streamflow cadence. In calculating accumulations, if any 2-minute interval within the accumulation window contained missing data, the entire accumulation window was flagged as missing data (assigned a value of NaN). The accumulated blocks of rainfall, along with FLASH CREST unit streamflow, were saved to be used as input data for ML model training and forecast generation.

2.3 High-Resolution Rapid-Refresh Model

The High-Resolution Rapid-Refresh model is a 3km convection-allowing model run operationally by the National Oceanic and Atmospheric Administration (NOAA). HRRR is initialized each hour and produces forecasts out to 18-48 hours of lead time depending on the run (Dowell et al. 2022). HRRR forecast fields were extracted from a publicly available NOAA dataset archived on Amazon Web Services (AWS).

For each HRRR run time from 00 UTC 1 May to 23 UTC 30 September, 2021-2025, a set of meteorological variables were extracted with the intention of providing the ML model relevant information regarding the prevailing synoptic and mesoscale environments. These variables included wind components, temperature, and dewpoint at multiple vertical levels (near surface, 850 hPa, 500 hPa), as well as precipitable water (PWAT) and surface-layer convective available

potential energy (CAPE). A mask to isolate CONUS with a border buffer was also applied similar to Chapter 2.2.

Since the HRRR initializes at least one hour prior to the decision times within an hour, all non-QPF atmospheric predictor fields were extracted from forecast hour (FH) 1 of the most recent run. These fields were valid for the hour block that contains the current decision time (t = 0 for the ML model). For models incorporating future precipitation, QPF was taken from later HRRR forecast hours. QPF at FH 2 represents accumulation from FH 1 (the start of the current hour) to FH 2 (the end of the current hour), corresponding to rainfall during the first hour of the ML forecast period. QPF at FH 3 represents accumulation during the second hour of the ML forecast period. The data used to train the ML model was thus analogous to that which would be available to an operational forecaster when making a nowcasting decision.

2.4 Land Cover

Because land cover can strongly impact flash flood potential, the ML model was provided with information about the spatial distribution of land use within its domain. USGS publishes the Annual National Land Cover Database (NLCD) (United States Geological Survey 2025). This database divides the United States into 16 different categories of land cover type. For simplification, categories likely to have similar impacts on flash flood potential were merged according to Table 1. This data is also much higher resolution than the HRRR grid, and category information needed to be preserved during interpolation. Therefore, within each HRRR grid cell, NLCD grid cells were accumulated by type. The category with the highest number of NLCD grid cells became the HRRR grid cell. For consistency, the CONUS mask was also applied.

NLCD Category Simplification	
Original	**Simplified**
Developed, Open Space	Developed
Developed, Low Intensity	
Developed, Medium Intensity	
Developed, High Intensity	
Grassland/Herbaceous	Grassland/Crops
Pasture/Hay	
Cultivated Crops	
Deciduous Forest	Forest
Evergreen Forest	
Mixed Forest	
Barren Land	Barren/Arid
Shrub/Scrub	
Woody Wetlands	Wetlands
Emergent Herbaceous Wetlands	
Perennial Ice and Snow	Perennial Ice and Snow
Open Water	Open Water

Table 1. Original NLCD categories compared to the more general category used as predictors for machine learning training.

2.5 Summary of Data

In summary, the complete set of 2D fields used to train the ML model are listed below in Table 2. This includes data from MRMS, FLASH, HRRR, and USGS land cover, all interpolated to the HRRR grid where necessary.

Model Inputs		
FLASH CREST Maximum Unit Streamflow (10 min, MRMS)	0-1, 1-3, 3-6, 6-12, 12-24 Hour QPE (10 min, MRMS)	Precipitation Rate (10 min, MRMS)
10m u-wind, v-wind (1 hour, HRRR)	2m temperature, dew point (1 hour, HRRR)	Precipitable water (1 hour, HRRR)
Convective available potential energy (1 hour, HRRR)	500mb/850mb u-wind, v-wind, temperature, dewpoint (1 hour, HRRR)	FH 2, 3 QPF (1 hour, HRRR)
Land Cover (Static, USGS)	Hour Fraction	**NWS flash flood warnings (10 min, NWS, IEM Database)**

Table 2. The features used in training the ML model. Parentheticals are the time steps of the data and sources of the data. Label dataset is in bold. Flash flood warning plots were generated in 10-minute increments but were combined into hourly blocks for some models described in Chapter 3.3.

Chapter 3: Methods

3.1 Model Selection

The prediction problem addressed in this study requires estimation of spatially distributed probabilities of FFW issuance on a fixed grid. At each decision time, the model must ingest multiple gridded predictor fields and produce a two-dimensional probability field. Unlike classification problems that predict a single scalar outcome, this task requires a model architecture capable of preserving spatial structure throughout the prediction process. With these considerations in mind, a U-Net convolutional neural network architecture was selected for this research.

The selection of this architecture was guided by two primary ideas. First, prior unpublished research performed by the Center for Analysis and Prediction of Storms (CAPS) has demonstrated the suitability of U-Net models for spatial prediction tasks using high-resolution meteorological data. Previous work outside of CAPS, such as Li et al. (2024), has also shown promise in U-Net forecast inference from radar data. Second, the structure of the prediction problem naturally aligns with the image to image learning framework used by U-Nets (Chase et al. 2023).

The U-Net architecture extends conventional CNNs by incorporating an encoder-decoder structure with skip connections between corresponding resolution levels. Figure 1 is a graphical representation of this process (reprinted from Fig. 6 of Chase et al. 2023). The left branch of the U, often referred to as the encoder pathway, progressively reduces image resolution through the convolutional layers and pooling. The bottom of the U transitions into a decoder pathway, the branch on the right. In this branch, feature maps are upsampled back to the original grid

resolution and passed through additional convolutional layers. After each upsampling step, the reconstructed feature maps are concatenated with feature maps of the corresponding resolution level of the encoding pathway. These skip connections help restore fine scale spatial detail that may have been reduced during pooling and allow the model to combine large scale context with localized structure.

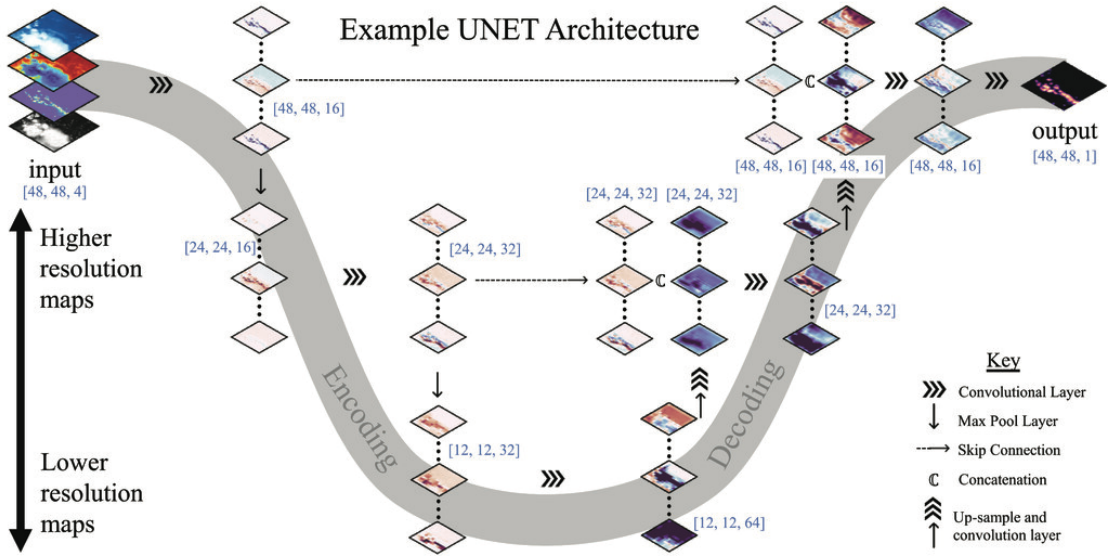

Figure 1. An example of a U-Net architecture's encoding and decoding structure with skip connections (Chase et al. 2023).

Learning occurs through the convolutional filters within each layer. Each filter consists of a set of trainable weights that interact with small neighborhoods of the input field, allowing the network to detect spatial patterns such as gradients, boundaries, or structures. During training, these weights are iteratively adjusted to minimize the chosen loss function. In effect, the model learns what combinations/patterns of rainfall intensity, hydrologic response, and environmental conditions most strongly correspond to FFW regions. The layering strategy in the encoder branch captures the impact of spatial features and variations at different scales and learns to relate them to the likelihood of FFW issuance. This allows the model to identify general regions of rainfall as

well as smaller scale high-intensity convection causing flash flooding, for example. Through this hierarchal process, the network develops an internal representation of patterns that are predictive of warning issuance.

3.2 Patching and Training Data Selection

Model training data were constructed using a patch-based sampling approach. Predictor and label fields were defined in 10-minute intervals spanning 00 UTC 1 May to 23 UTC 30 September, 2021-2023. For each training sample, a single time was selected at random from this timeline, the full set of input fields were opened once for that timestamp, and then multiple spatial patches were extracted from random locations across the domain.

Each patch contains 128 by 128 pixels, which on a 3 km grid covers an area of approximately 384 km by 384 km. For every selected decision time, 200 random patch centers were drawn within the valid spatial domain. Each patch was inspected for the quality control and class-based constraints below. This process was repeated until 300,000 patches were gathered to complete the training dataset.

Because FFWs are rare events, their appearance in the training data would be very scarce without preferentially sampling them. Class balance was thus central in the patch selection strategy. Therefore, to control class balance, each patch was assigned to one of three categories based on the warning label and rainfall presence in the patch:

1) **WR:** Warning with rain
2) **NWR:** No warning with rain
3) **NN:** No warning with no rain

The target data composition was 40% WR, 45% NWR, and 15% NN, providing the ML model with a balanced mix of examples containing rainfall with and without FFWs as well as no rainfall. With 300,000 total patches, this corresponds to 120,000 WR, 135,000 NWR, and 45,000 NN samples. Data were written into uncompressed TFRecord shards of 10,000 patches each, resulting in 30 shards. Shards were filled independently, meaning patches were written directly into the active shard until class quotas were met, with no queueing or carryover. Once a shard's quota for a category was reached, additional candidate patches from that category were skipped, ensuring each shard had a consistent class mixture.

Rainfall presence was determined using the MRMS QPE, with a patch considered rainy if any QPE field exceeded 1.0 mm within the patch. Quality controls were also applied during sampling. Candidate patches were rejected if more than 50% of the QPE values within the patch were missing, which occurred when gaps were present in the MRMS data (see Chapter 2.2). Additionally, patches were rejected if the overall fraction of missing values across all predictor channels exceeded 30%. This prevented patches dominated by incomplete or missing input data from entering the training dataset.

To approximate an operational setting, all predictor fields were time aligned relative to the decision time (t=0). For each 10-minute increment, only information that would have been available at or before that moment was used. Observational inputs from MRMS were treated as delayed by up to 10 minutes (t=-10 minutes), meaning data were sampled from the most recent analysis time preceding the decision time. HRRR predictor fields were selected from the most recent completed model run available at that time (t =-1 hour rounded down, see Chapter 2.3). These were paired with the FFW label dataset valid at the decision time.

Finally, a simple time awareness predictor was added to every patch. Since model and observation data will overlap within hour blocks, a constant channel representing the fractional hour of the decision time was included. This channel was defined as the fraction of minutes past the top of the hour scaled down to [0,1). The purpose of this is to give the model a sense of time within the hour. At the beginning of the hour, it should rely more on forecast information, and at the end of the hour, it should rely more on observations. Predictor channels were stored in a consistent order, with the mean and standard deviation computed incrementally across accepted patches for standardized model training and inference.

An independent validation dataset was constructed using the same patch-based framework described above but applied to the subsequent warm season spanning 1 May to 30 September 2024. All preprocessing steps, operational timing rules, predictor channels, quality controls, and patch extraction procedures were kept consistent with the training dataset. However, unlike the training dataset, patches for validation were sampled randomly without class-based balancing in order to better reflect the natural distribution of warning and non-warning cases. The validation dataset was also smaller, consisting of 80,000 patches rather than 300,000. No data from the 2021-2023 training period were included. This dataset was used solely during model training for hyperparameter tuning and early stopping, while Chapter 3.4 details the use of 2025 warm season data for testing and verification of model performance.

3.3 ML Model Configurations

Three model configurations were trained using progressively more forecast precipitation information. The No-QPF configuration used only predictors operationally available at the

decision time (MRMS and FLASH fields, along with HRRR environmental fields). The target label for this configuration was whether an FFW was active within the following 10-minute block.

The main challenge with incorporating forecast information is that the HRRR only models QPF in hourly blocks. Every forecast hour, on the hour, the HRRR QPF is calculated for the entire hour block. However, the other data has a much higher temporal resolution and could offer sub-hourly information regardless of HRRR update times. Therefore, a 10-minute cadence is maintained for decision times by the ML model, but the HRRR data was handled with the method described below.

The +1h-QPF configuration added HRRR 1-hour accumulated precipitation for the hour block containing the decision time. Operationally, this accumulation is obtained from the most recent available HRRR run, which would be an hour behind the current hour block. Therefore, QPF within the current hour block corresponds to accumulation between FHs 1 and 2, which is contained within FH2. The target label was expanded to include warnings valid at any time within that hour. As an example, if the decision time is 0630 UTC, then the 0500 UTC HRRR is the latest available run. 0500 UTC becomes FH0, 0600 UTC becomes FH1, and 0700 UTC becomes FH2. To get QPF for the current hour block, so in this case between 0600 UTC and 0700 UTC, FH2 is used.

The +2h-QPF configuration added HRRR QPF for the subsequent UTC hour (e.g., 07-08 UTC) block, corresponding to the 1-hour QPF between FHs 2 and 3. The target label was expanded accordingly to include warnings occurring within both the decision-time hour and the following hour.

3.4 Model Training and Optimization

3.4.1 Training Framework

All models were implemented in TensorFlow using the keras_unet_collection (Yingkai (Kyle) Sha 2022) package. The architecture consisted of a U-Net configured for binary segmentation with a single output channel representing the pixel probabilities of FFW issuance. Convolutional layers used ReLU activation functions, and the final output layer applied a sigmoid activation to produce probabilities bounded between 0 and 1. Batch normalization was applied within convolutional blocks to improve training stability. The network depth, maximum channel width, and filter progression strategy were treated as tunable hyperparameters (Chapter 3.4.3). Prior to training, each predictor channel was standardized using the training dataset means and standard deviations to put variables on a common scale and improve training stability, even though the underlying distributions may not be normal. During data loading, NaN values were replaced with the corresponding channel mean before normalization to avoid errors.

Training was performed using the Adam optimizer. The learning rate was treated as a tunable hyperparameter (Chapter 3.4.3). Models were trained using a distributed strategy when multiple GPUs were available, allowing synchronous gradient updates across devices. Training data were streamed from TFRecord shards using the TensorFlow tf.data pipeline. The number of steps per epoch was determined from the total number of samples across all shards to ensure full coverage of the dataset during each epoch.

Model performance was monitored on the independent validation dataset at the end of each epoch using precision-recall area under the curve (PR-AUC). PR-AUC was selected because FFWs occupy only a small fraction of the spatial domain, resulting in a strongly

imbalanced classification problem. Unlike accuracy or ROC-AUC, precision and recall exclude true negatives and therefore provide a more informative assessment of rare-event prediction. PR-AUC summarizes performance across all probability thresholds, favoring models that maintain both high precision and high recall. Early stopping with a patience of 10 epochs was applied to prevent overfitting. The final retained model weights correspond to the epoch achieving the highest validation PR-AUC.

3.4.2 Loss Function

Training used a composite loss function (Eq. 1) combining binary cross entropy (BCE) and Dice loss (Xu et al. 2023). The weighting parameter α was tuned during hyperparameter optimization to balance sensitivity to rare warning pixels with overall segmentation performance:

$$\mathcal{L} = \alpha * BCE + (1 - \alpha)(1 - Dice) \quad (1)$$

BCE optimizes pixel-wise classification accuracy by penalizing each pixel independently. In contrast, Dice loss evaluates the global overlap between predicted and observed regions using a normalized sum over all pixels in a patch. Since Dice operates on aggregated region-level agreement rather than independent pixel errors, it encourages spatially coherent and contiguous warning structures rather than fragmented pixel predictions. The composite BCE-Dice formulation therefore combines strong pixel-level discrimination with an explicit region-overlap constraint, improving the structural consistency of the predicted warning areas while preserving probabilities.

3.4.3 Hyperparameter Optimization

For each dataset configuration described in Section 3.3, model training included optimization of hyperparameters using the Hyperband algorithm (Li et al. 2017) implemented in the KerasTuner python package. Hyperband was selected because it efficiently explores a broad hyperparameter space while allocating progressively more resources to higher performing configurations. Model performance during hyperparameter tuning was evaluated using pixel-level PR-AUC of the validation dataset.

A full list of the hyperparameter search space is as follows:

- Network depth (3 to 5 encoder decoder levels)
- Maximum convolutional channel width (64, 128, 256)
- Filter progression strategy (constant channels or doubling with depth)
- Learning rate for Adam optimizer (1e-5 to 3e-3)
- Loss weighting parameter α controlling the balance between binary cross entropy and Dice loss (0.2 to 0.8)

Each trial was dynamically trained for a number of epochs determined by the Hyperband algorithm. To reduce wall clock time during tuning, the number of training and validation steps per epoch were capped. This ensured consistent comparisons between configurations while accelerating the parameter space search. Table 3 shows the best results of the hyperparameter search.

Best-Performing Models from Hyperparameter Search		
+2h-QPF	**+1h-QPF**	**No-QPF**
Depth: 3 Max channels: 256 Layer strategy: constant BCE Dice α: 0.8 Learning rate: 0.0003 PR-AUC Score: 0.2411	Depth: 3 Max channels: 128 Layer strategy: constant BCE Dice α: 0.8 Learning rate: 0.0003 PR-AUC Score: 0.2529	Depth: 4 Max channels: 256 Layer strategy: constant BCE Dice α: 0.8 Learning rate: 0.0005 PR-AUC Score: 0.2881

Table 3. A summary of the hyperparameter space found to be highest scoring in PR-AUC by Hyperband.

3.4.4 Retraining Final Models

After Hyperband tuning, the highest performing hyperparameter set for each model configuration was selected. The models were then fully retrained using the procedure outlined in Chapter 3.4.1. The independent 2024 validation dataset was used exclusively for model selection and early stopping, while 2025 warm season data were reserved for final performance evaluation and verification.

3.5 Validation

To evaluate model performance in a realistic forecasting setting, an independent test dataset was aggregated for the 2025 warm season spanning 1 May to 30 September 2025. For

each decision time on a 10-minute interval, predictor fields were assembled using the same operational timing assumptions in Chapters 2.3 and 3.2. These fields were the full 1059 by 1799 CONUS grids. Static land cover fields and the fractional hour predictor were added consistent with the training configuration. Predictor channels were normalized using the same mean and standard deviation statistics computed from the training dataset. If required input data were missing for a given time, that forecast was skipped and recorded as such.

Instead of randomly sampling patches, the full domain was decomposed into overlapping 128- by 128-pixel patches using a stride of 64 grid points. Each patch was passed through the trained U-Net model on the GPU to generate pixelwise probability predictions. To reconstruct a seamless full grid forecast, overlapping patches were combined using a Hann window weighted overlap add procedure (Pielawski and Wählby 2020). This approach reduces edge artifacts introduced by patch-based inference and ensures smooth transitions between adjacent regions. After stitching, predictions were masked to the CONUS domain. These files constitute the independent 2025 test dataset used for evaluation and verification in future chapters.

Chapter 4: Model Verification

4.1 CONUS

Figure 2 presents precision-recall (PR) curves and corresponding F1 scores for the operational model configurations during the 2025 warm season across CONUS. Each configuration is evaluated against FFW within the lead times considered by each model, referred to here as the native warning label definition. The No-QPF model is verified against warnings active within the next 10 minutes. The +1h-QPF model is verified against warnings active within

the fixed hourly block containing the forecast valid time, and the +2h-QPF model is verified against warnings active within that same hourly block plus the following hour block. Thus, the QPF-based models are evaluated using hour-aligned label windows rather than rolling 1- or 2-hour periods.

In the precision-recall curves, all model configurations demonstrate meaningful predictive skill across a range of probability thresholds. The No-QPF model consistently performs best, maintaining higher precision for all recall levels. As forecast lead time increases to +1 hour and +2 hours, the PR curves indicate reduced precision for a given recall and a decrease in the maximum attainable precision.

The F1 scores further illustrate threshold-dependent behavior. For all lead times, F1 is maximized over a probability of around 0.5-0.6, suggesting that the models use their probabilities relatively similarly. However, peak F1 decreases with forecast lead time, with the +2h-QPF model exhibiting the largest reduction in skill. Higher probability thresholds favor precision at the expense of recall, while lower thresholds increase event detection but introduce more false alarms. In operational use, a forecaster would need to consider the tradeoff between missed events and over-warning.

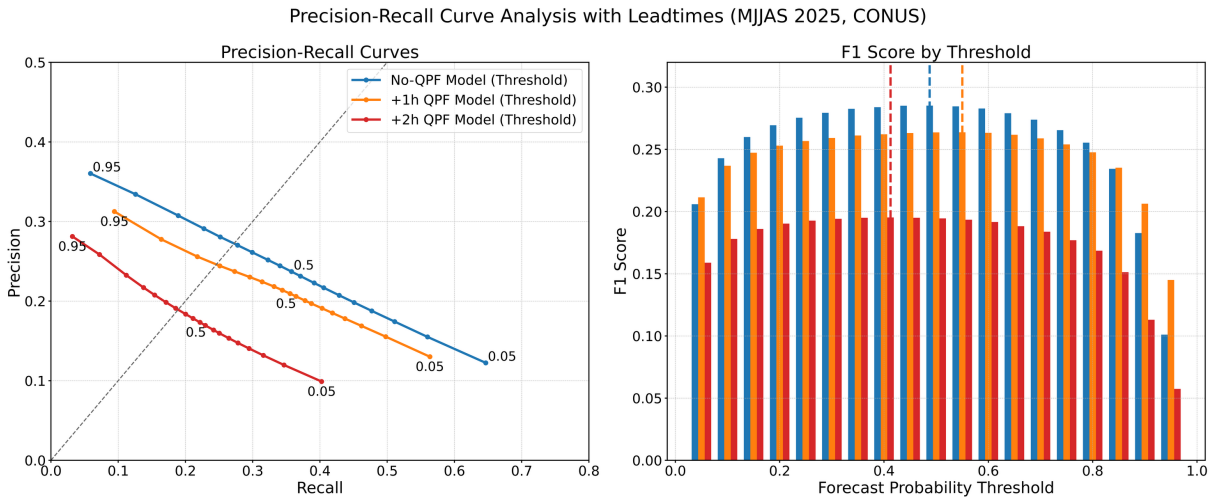

Figure 2. Precision-Recall (left) and F1 score by probability threshold (right) for the CONUS domain during the 2025 test period. Performance is shown for the No-QPF (blue), +1h-QPF (orange), and +2h-QPF (red) models. In the precision-recall space, dots denote probability thresholds in increments of 0.05. The dashed diagonal line represents a 1:1 precision-recall ratio. Values on the precision-recall plot can vary from 0 to 1, but the plot is zoomed in to show more detail. The vertical dashed lines in the right plot denote the thresholds for maximums in F1 scores.

The distribution of warning-relative lead times, when newly warned pixels first exceed selected probability thresholds relative to the official FFW issuance time, is plotted in Figure 3. With the QPF models, each model was trained on hourly blocks of warning information that align with available HRRR QPF information. Therefore, a forecast from the +2h-QPF model will be valid for the current hour block (if it's 0320 UTC, that would be 0400 UTC to 0500 UTC) and the next hour block (0500 UTC to 0600 UTC). When the hour changes over, the forecast then becomes valid for the next hour blocks. The +1h-QPF model would only be valid for the current

hour block. Warning-relative lead time is defined as the number of minutes between when a pixel first surpasses a probability threshold and when the warning was issued, with negative values indicating that the ML model exceeded the threshold before the issued warning(s). It is important to note that Figure 3 only includes pixels that exceeded the specified probability threshold at some point during the evaluation window. Pixels that never crossed a given threshold are not represented in the warning-relative lead time distributions. Approximately 50% of all newly warned pixels did not cross T = 50% for the No-QPF model, and ~40% did not cross T = 50% for the +1h-QPF and +2h-QPF models. Those values increase with increasing thresholds to 62-77% at T = 90%.

At the 50% threshold (T = 50%) (Figure 3a), all configurations exhibit broad distributions centered at various distances from 0 minutes. The No-QPF model has a peak in the number of pixels at +10 minutes (or 10 minutes after the warnings were issued). The +1h-QPF model shows a peak at 0 minutes, and the +2h-QPF model peaks at -20 minutes. This trend in earlier probability thresholds is more apparent when plotting a cumulative distribution function (CDF) (Figure 3d), where the accumulated number of preceding pixels are divided by the total number of pixels. When plotting the CDFs for each model, the +2h-QPF model shows a significantly higher cumulative fraction of probability exceedance pixels than the No-QPF model alone. For example, the +2h-QPF model has 47% of its pixels over 50% probability at 30 minutes before warning issuance, while the No-QPF model has only 24% of 50%+ pixels. This 23% difference in pixel thresholds indicates that the models which incorporate HRRR QPF information, especially the +2h-QPF model, appear to be gaining valuable information from forecasted rainfall, or at least are confident earlier than the No-QPF model.

With increasing probability thresholds (Figures 3b, 3c), the histograms of all three models begin to converge toward 0 minutes before warning issuance (i.e., the ML model identifies events at approximately the same time as warnings are issued). There is also a decrease in the peak number of pixels with all three histograms, indicating not as many pixels cross the higher warning thresholds. A decrease in the peak warning-relative lead time and number of pixels at those peaks indicates the models are becoming more uncertain. The CDFs reflect this reduced confidence through the decrease in space between the three models' lines (Figures 3e, 3f). Specifically, the models attempting to predict further into the future are less confident, likely due at least in part to the forecast error inherent to QPF. Using the same 30-minute difference from the previous example at T = 90% instead of T = 50%, the difference between the No-QPF and +2h-QPF models' CDFs reduces to 8%.

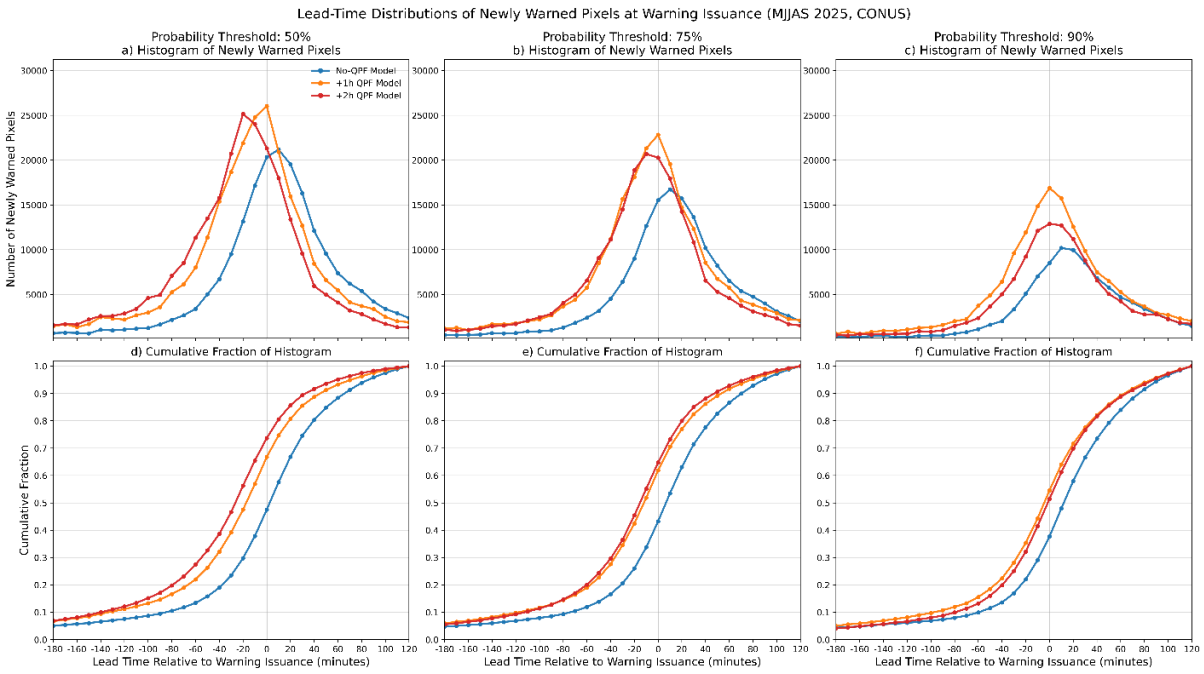

Figure 3. Lead time distributions of newly warned pixels relative to flash flood warning issuance time during the 2025 test period over CONUS. Columns correspond to probability thresholds of 50%, 75%, and 90%. **(a, b, c)** Top panels show histograms of lead time (minutes), while **(d, e, f)** bottom panels display the corresponding CDFs. Results are shown for the No-QPF (blue), +1h-QPF (orange), and +2h-QPF (red) models. Negative values indicate pixels that meet a threshold before the warning was issued, zero is warning issuance, and positive values are after.

An important caveat to Figure 3 is presented in the reliability diagram in Figure 4. All three models consistently over-forecast across most probability bins, as the conditional observed relative frequency remains below the 1:1 reference line. The persistence of over-forecasting across configurations suggests that the models tend to assign probabilities that are too large relative to the observed frequency of warnings. This behavior is most pronounced at higher forecast probabilities and as additional HRRR QPF information is added. This aligns with the

trend in QPF models providing more lead time and could mean that benefits in lead time are coming from more bullish predictions, not future insight. Several mechanisms may contribute to the general overconfidence of the ML models. These possible mechanisms are explored further in the following sections.

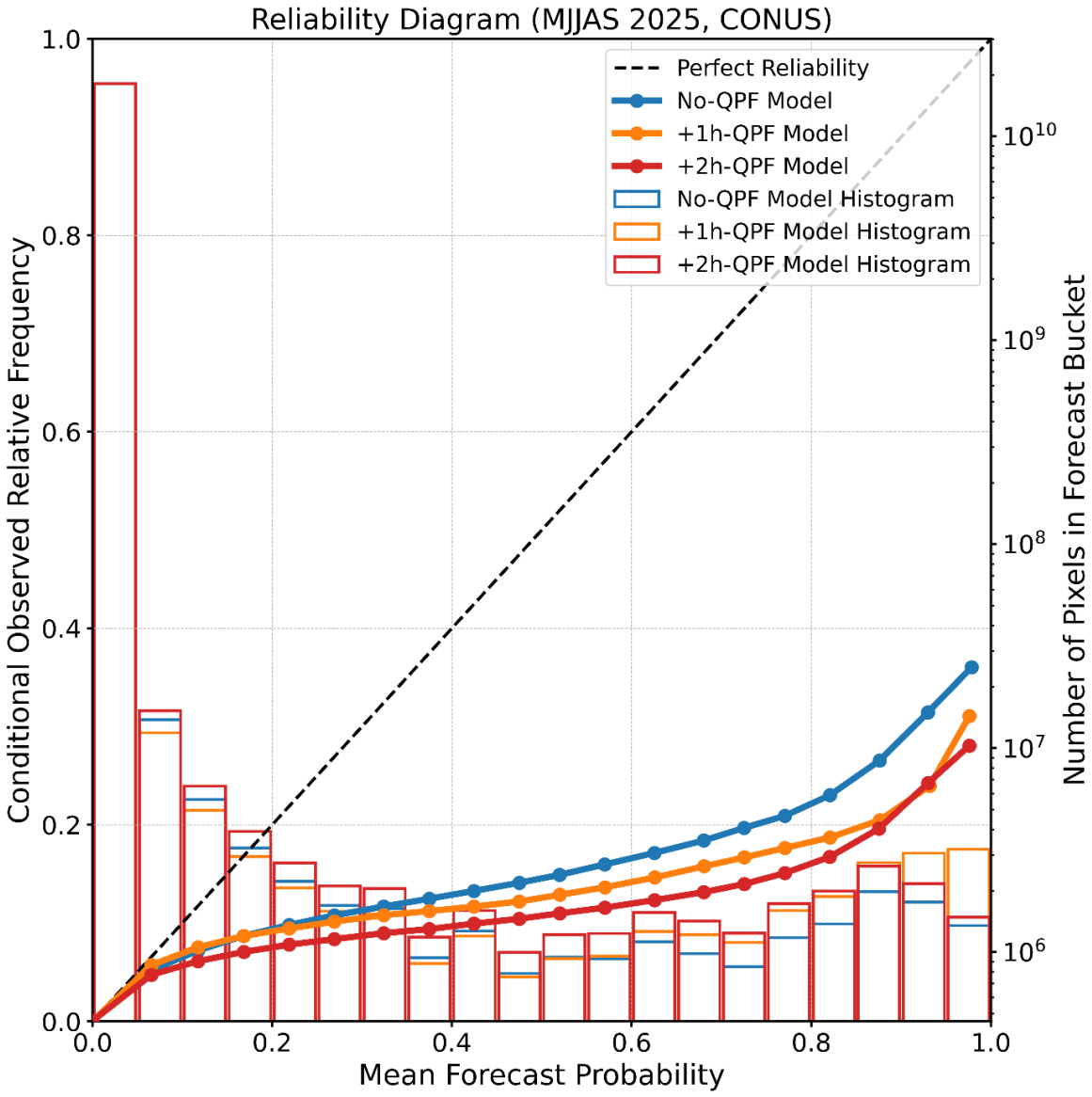

Figure 4. Reliability diagram for the MJJAS 2025 test period over CONUS. Curves show conditional observed relative frequency as a function of mean forecast probability for the No-

QPF (blue), +1h-QPF (orange), and +2h-QPF (red) models. Each bar uses the same model colors and represents the number of pixels in each mean forecast probability bucket. The dashed diagonal line represents perfect reliability.

The false alarm ratio (FAR) as a function of forecast probability threshold is shown in Figure 5 for each model configuration during the MJJAS 2025 test period over CONUS. FAR represents the fraction of forecast pixels exceeding a given probability threshold that did not correspond to an active FFW. This metric provides a complementary perspective to the precision-recall analysis and has important operational implications.

Across all configurations, FAR decreases steadily as the probability threshold increases. At lower thresholds, the models predict FFW issuance across a larger number of pixels, resulting in higher FARs. As the threshold increases, the models become more selective in predicting potential warnings, reducing the fraction of detections that correspond to non-warning pixels. This behavior reflects the same tradeoff between detection and false alarms seen in the precision-recall curves of Figure 2, where lower thresholds favor higher recall at the expense of precision.

Differences between model configurations are also apparent. The No-QPF model consistently exhibits the lowest FAR across nearly all probability thresholds, indicating that a larger fraction of its detections corresponds to active warning regions. In contrast, the models incorporating QPF show progressively higher FAR values, with the +2h-QPF configuration producing the largest ratios. This behavior is consistent with the degradation in precision observed in Figure 2 and reflects the increasing uncertainty associated with relying on a combination of QPE and QPF rather than purely on QPE. At higher probability thresholds, the

FAR values of all configurations decrease noticeably, indicating that the most confident predictions from each model are more likely to correspond to active warning regions.

Despite the decrease in FAR at higher thresholds, a substantial fraction of forecast pixels remains false alarms on the CONUS scale. This indicates that even the most confident predictions of FFW issuance included substantial portions outside active warnings when evaluated across the full national domain. However, it is important to remember that the validation results up until this point are strictly pixel-by-pixel with no buffer. If probabilities are outside of a warning by even one pixel, that will count as a false alarm, which could be inflating FAR in Figure 5.

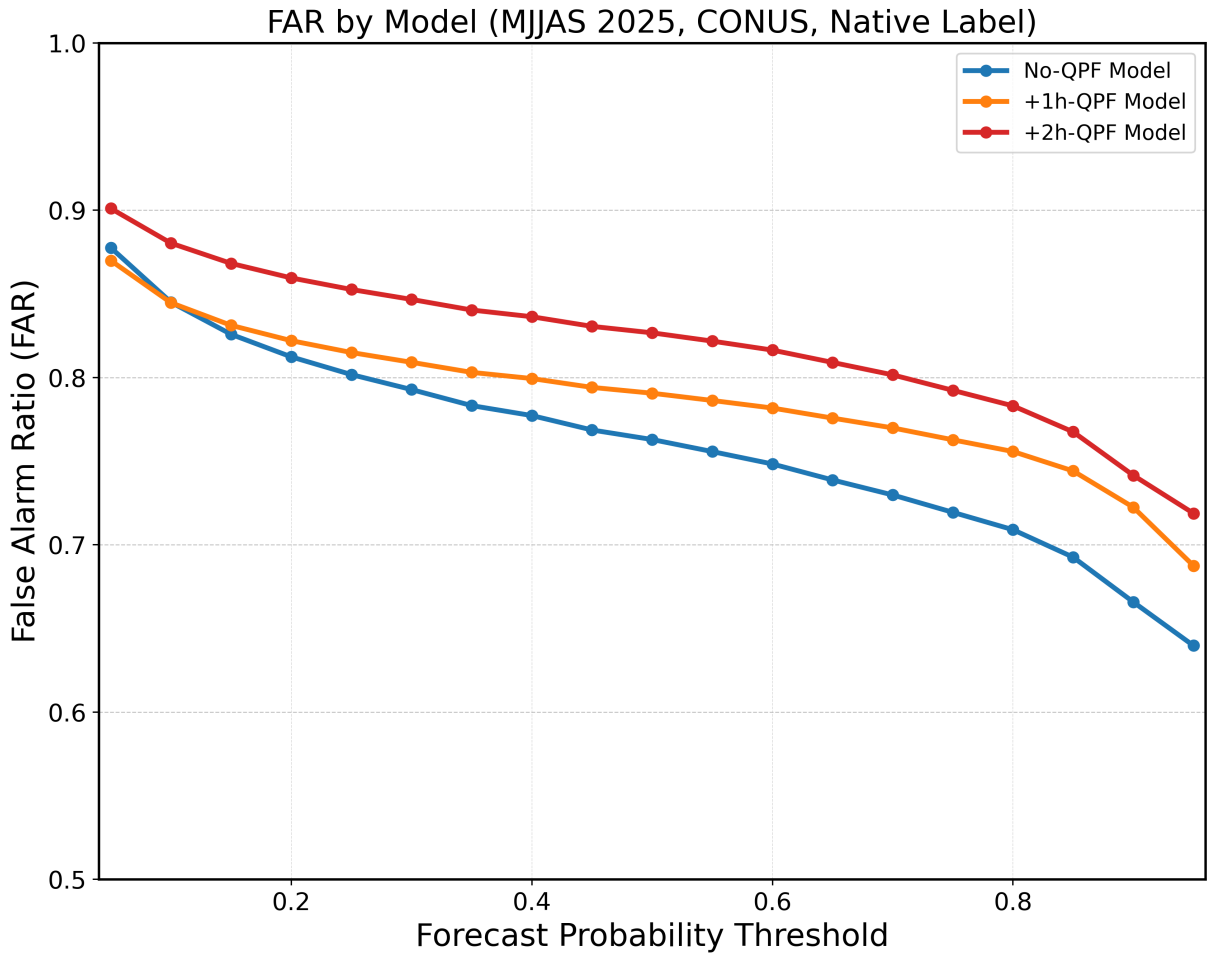

Figure 5. False alarm ratio (FAR) versus forecast probability threshold for the MJJAS 2025 test period over CONUS. Curves are shown for the No-QPF model (blue), +1h-QPF model (orange), and +2h-QPF model (red). FAR can range from 0 to 1, but the plot is zoomed in on the upper portion of this range to better show detail.

The CONUS-wide statistics presented thus far represent an aggregate of conditions across a large and diverse domain. Differences in observational coverage, hydrologic response, convective environments, warning issuance policies among NWS offices, and other factors can influence how flash flood signals are represented in both the predictor fields and warning dataset.

As a result, the national metrics shown in the preceding figures may mask important regional differences in model performance. Strict pixel-based verification can penalize small spatial offsets between forecast probabilities and warning polygons, potentially inflating apparent false alarms. Finally, while the verification in this study only considers FFW, operational hydrologic messaging is a challenging and nuanced endeavor. An event which includes flash flooding might coincide with other types of flooding, and official hydrologic messaging could include a combination of FFW, flood warnings (FW), or flood advisories (FA) for the same event. Forecast probabilities associated with events for which NWS offices chose to issue FW or FA instead of FFW would therefore be counted as false alarms when evaluated strictly against the FFW dataset.

This combination of CONUS-wide aggregation, strict pixel-based verification, and the use of a single warning category may contribute to the general tendency for over-prediction in the ML models. The following sections of Chapter 4 examine regional variation in verification statistics and explore the sensitivity of these results to spatial tolerance. They also consider other operational NWS products used to communicate flooding, providing a more complete perspective on ML model behavior.

4.2 Regional Verification

The spatial distribution of observed FFW occurrences during the MJJAS 2025 evaluation period is plotted in Figure 6a. In addition to describing where warnings were most common, this panel shows the local relative event base rate. Warning activity is strongly concentrated across the central and eastern United States, with particularly high frequencies extending from the

southern Plains through the Midwest and into portions of the Mid-Atlantic, indicating that flash flooding (or at least the issuance of FFW) is more common in these regions. In contrast, substantially fewer warnings were issued across the northwestern United States and most of Florida.

The spatial distribution of the No-QPF model detections at T = 50% (Fig. 6b) broadly exhibits similar regional patterns. Higher predicted “yes” counts occurred over the same central and eastern regions where warnings were most common. The model also captures the relatively low frequency of warnings across the northwestern United States. However, notable differences are present in some regions. In particular, the model produces substantially more detections across Florida compared to the observed warning counts. This could be due to predictors that the ML model was not explicitly trained on, such as local geology or NWS office warning practices.

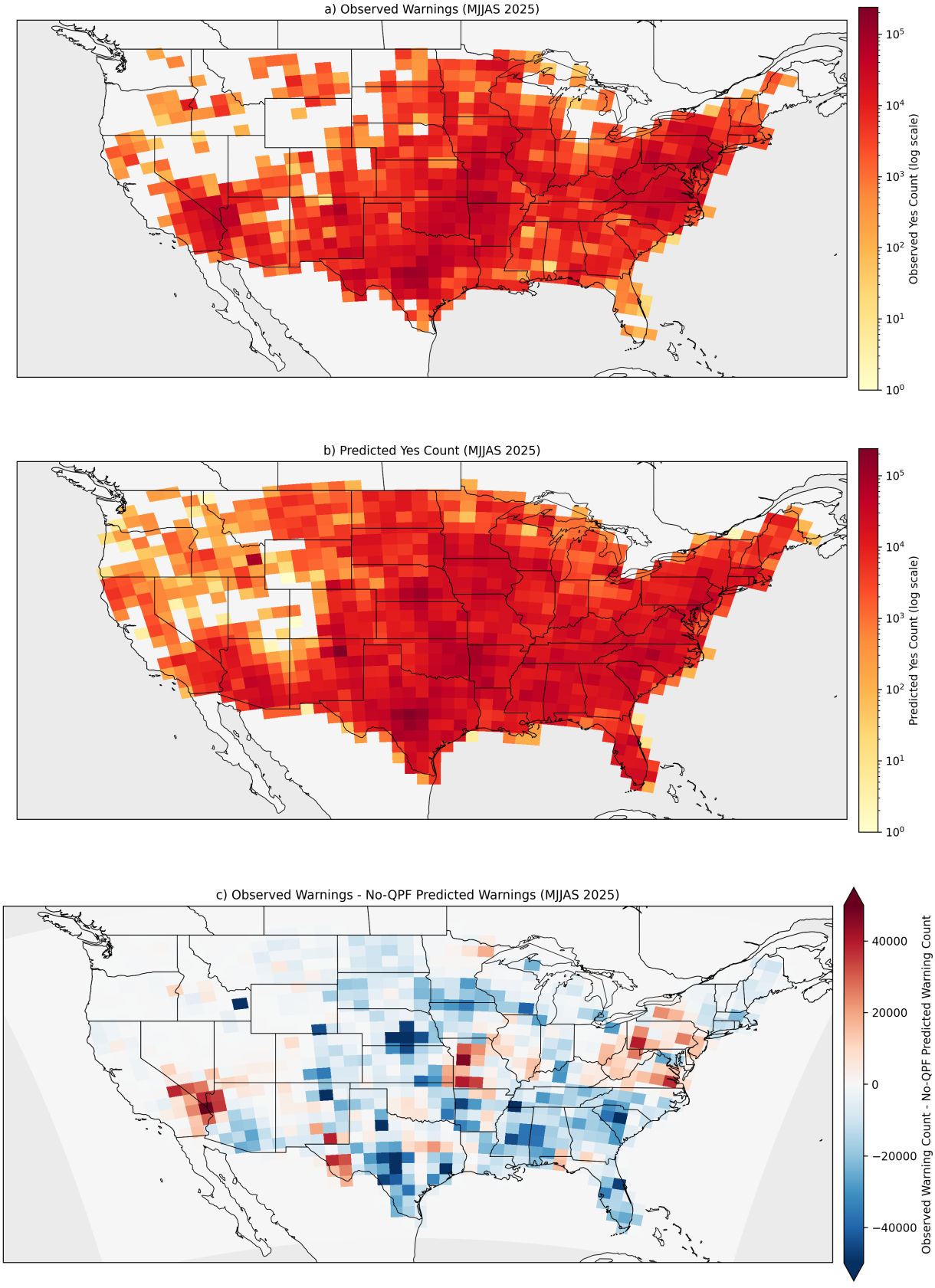
a) Observed Warnings (MJJAS 2025)
Observed Yes Count (log scale)
b) Predicted Yes Count (MJJAS 2025)
Predicted Yes Count (log scale)
c) Observed Warnings - No-QPF Predicted Warnings (MJJAS 2025)
Observed Warning Count - No-QPF Predicted Warning Count

Figure 6. Spatial distribution of warning pixels accumulated for each 10-minute decision-time increment during the MJJAS 2025 testing period. Counts represent the total number of grid cells meeting the specified condition at each 10-minute timestep, accumulated within 33x33 grid-cell neighborhoods across the CONUS domain for the full test period. **(a)** Observed flash flood warning counts displayed on a logarithmic scale **(b)** No-QPF model detections at a 50% probability threshold displayed on a logarithmic scale. **(c)** Observed flash flood warning counts (Figure 6a) minus the No-QPF predicted warning counts at the 50% threshold (Figure 6b).

The spatial distribution of forecast skill for the No-QPF model is shown in Figure 7a using F1. Model performance exhibits strong regional variability across the CONUS domain. The highest F1 values are concentrated across portions of the central United States, particularly from the southern Plains through parts of the Midwest, where F1 locally exceeds 0.35. Substantially lower F1 values are observed across much of the northwestern United States and Florida. This aligns with the patterns noted above in the frequency plots of Figure 6, as F1 is generally higher in regions where warnings were more common and the ML models correspondingly predicted more warnings. These regional differences are further highlighted in Figure 7b, which shows deviations of local F1 values from the national average. Among the potential contributors to this regional F1 pattern are the differences in the local relative event base rate shown in Figure 6a, as areas with more frequent observed warnings provide more opportunities for correct forecasts. Therefore, higher F1 across the central and eastern United States should not be interpreted as purely higher regional skill.

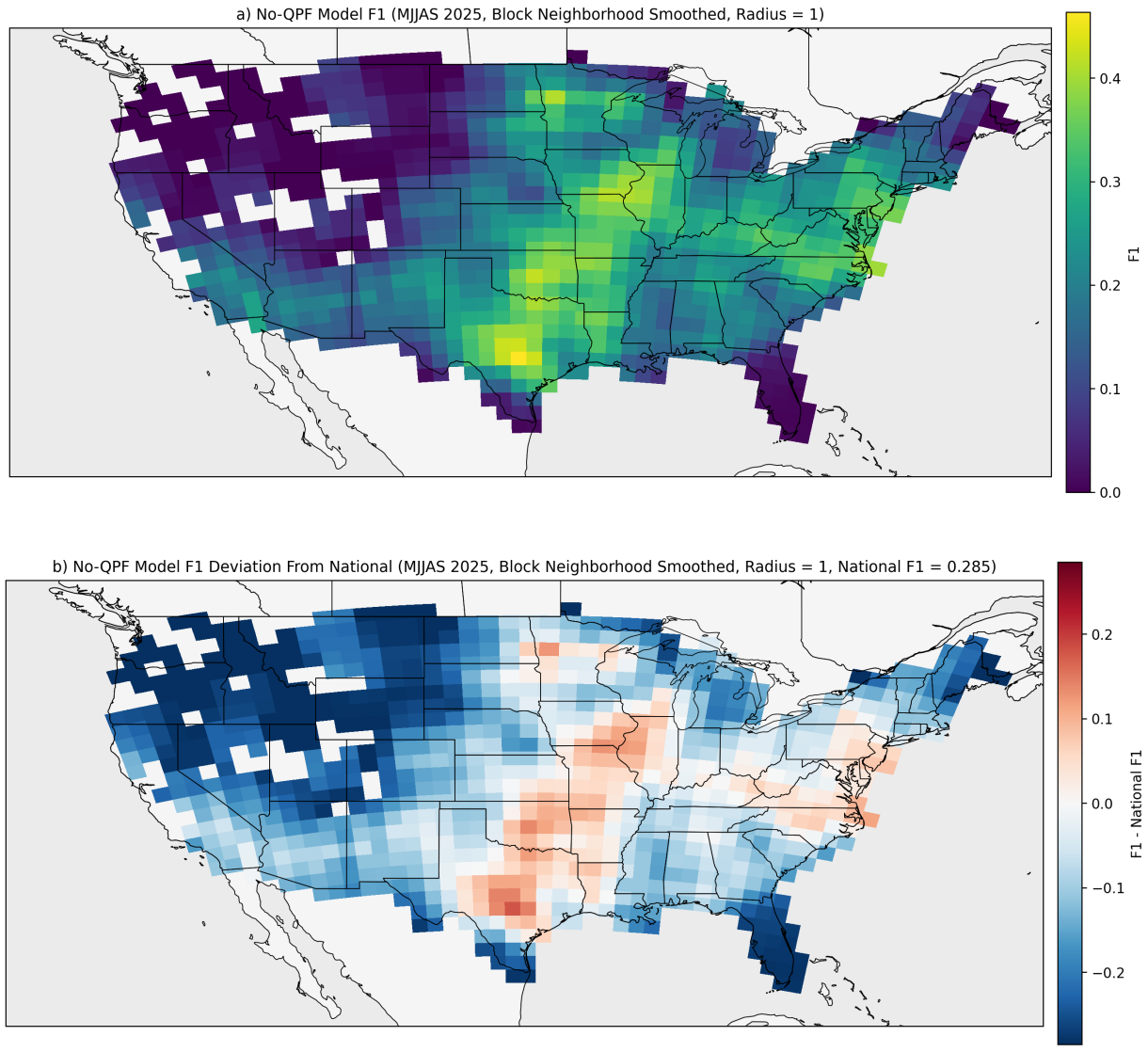

Figure 7. Spatial distribution of forecast skill and regional deviations from the national average during the MJJAS 2025 testing period for the No-QPF model. Verification statistics were calculated within 33x33 grid-cell neighborhoods at the 50% prediction threshold across the CONUS domain and subsequently smoothed using a 1-block radius neighborhood average. **(a)** F1 across the CONUS domain. **(b)** Deviation of local F1 values from the national F1 (National F1 = 0.285).

To further investigate the effects of base rate on F1, Figure 8 compares local F1 values with the observed warning counts in the smoothed 33x33 grid cell neighborhoods. A general positive relationship is evident, with higher observed warning counts tending to correspond to higher F1 values. This suggests that local relative base rate partially contributes to the regional F1 patterns shown in Figure 7. However, substantial variation remains around this overall trend, with a wide range of F1 values occurring at similar observed warning counts. This indicates that base rate alone does not fully explain the regional structure of forecast skill.

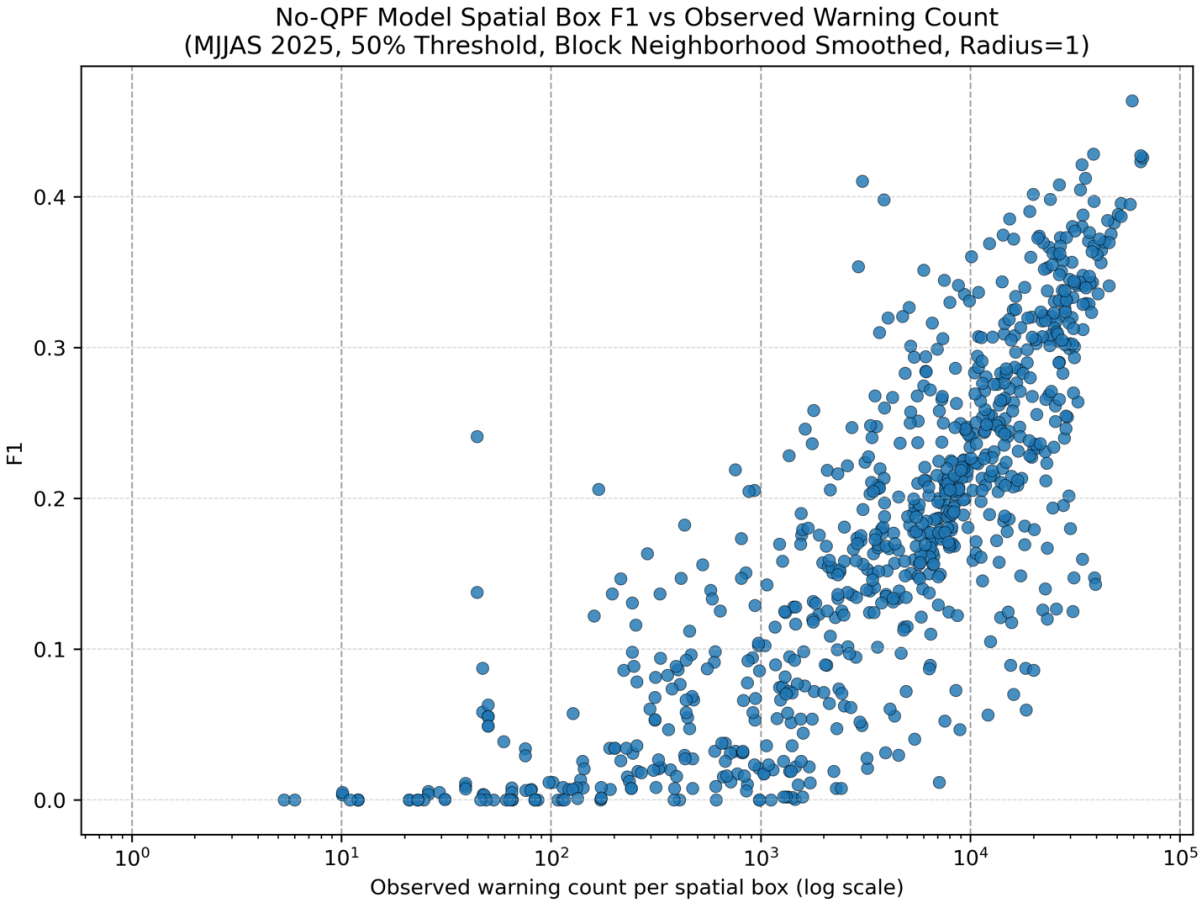

Figure 8. F1 at the 50% probability threshold as a function of observed warning count within each smoothed 33x33 grid-cell verification box during MJJAS 2025 over CONUS. The x-axis is shown on a logarithmic scale. Each point represents one spatial box.

The spatial distribution of FAR from the No-QPF model is shown in Figure 9. Consistent with the F1 patterns in Figure 7, FAR exhibits strong regional variability across the CONUS domain. Lower FAR values are concentrated across portions of the central and eastern United States, indicating that a larger fraction of model detections correspond to active warning regions in these areas. Substantially higher FAR values are observed across much of the northwestern United States and portions of Florida. These results further illustrate the strong regionality of model performance.

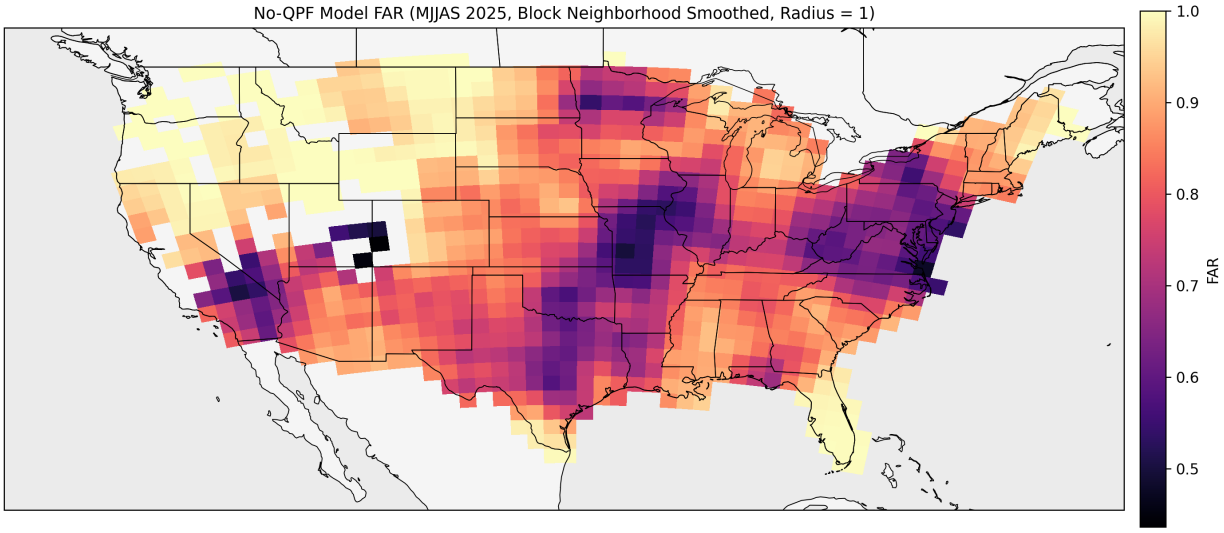

Figure 9. Spatial distribution of FAR during the MJJAS 2025 testing period for the No-QPF model. Verification statistics were calculated within 33x33 grid cell neighborhoods at the 50% prediction threshold across the CONUS domain and subsequently smoothed using a 1-block radius neighborhood average.

The spatial verification results for the +1h-QPF model are shown in Figure 10. Overall, the regional structure of forecast skill remains broadly similar to the No-QPF model. Figure 10a shows higher F1 values remain concentrated across portions of the central and eastern United States, while lower F1 values were found across the northwest United States and Florida. Also, compared to the No-QPF model, F1 generally decreases across most of CONUS. These reductions are modest in magnitude and do not substantially alter the spatial pattern of forecast skill.

The FAR for the +1h-QPF model is shown in Figure 10b. Compared to the No-QPF model, FAR increases across several regions, with notable increases over parts of Missouri and surrounding portions of the central United States. Outside of these areas, the spatial distribution of FAR remains broadly consistent with the patterns observed in Figure 9.

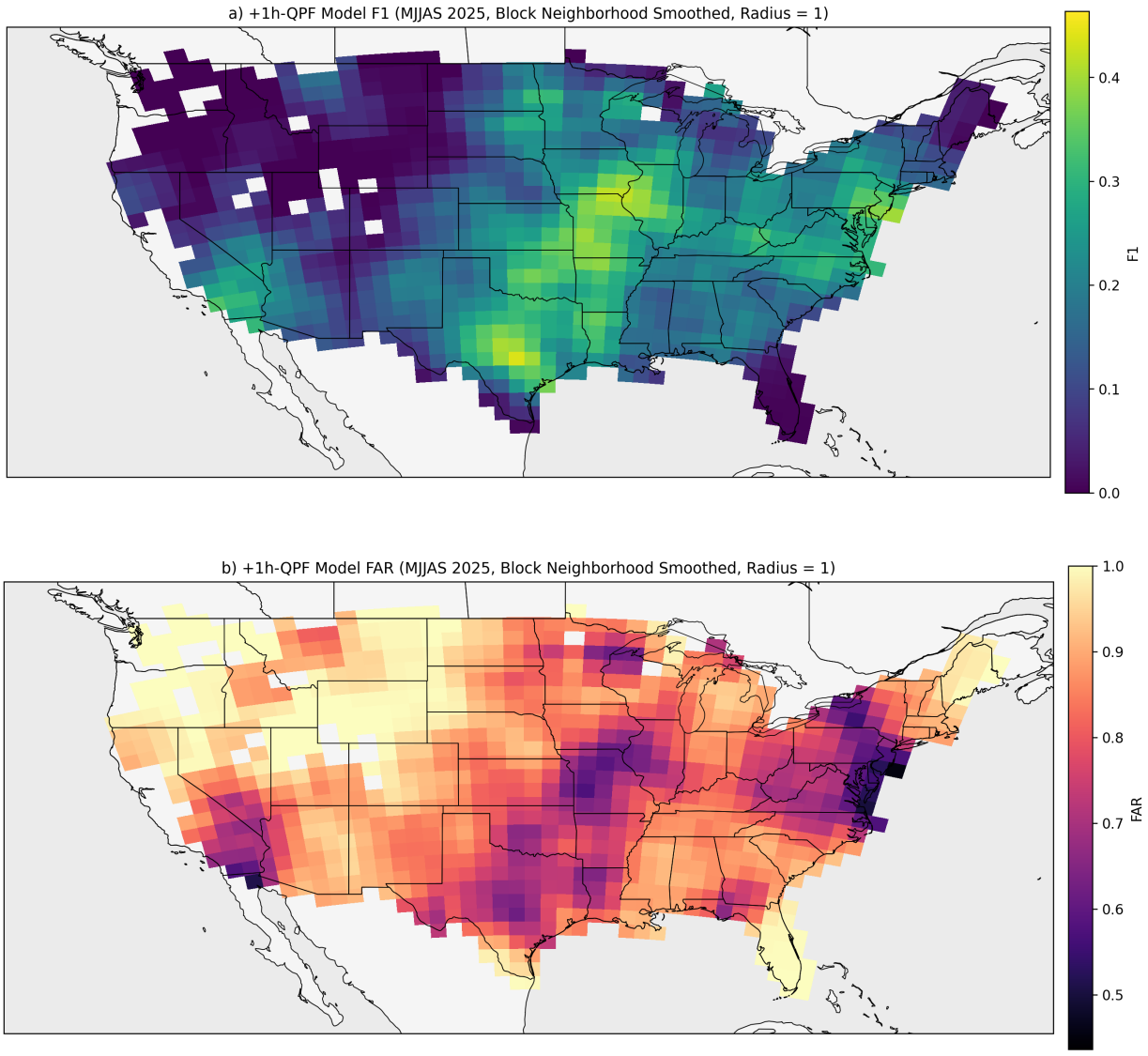

Figure 10. Spatial distribution of forecast skill and false alarm behavior during the MJJAS 2025 testing period for the +1h-QPF model. Verification statistics were calculated within 33x33 grid-cell neighborhoods at the 50% prediction threshold across the CONUS domain and subsequently smoothed using a 1-block radius neighborhood average. **(a)** F1 across the CONUS domain. **(b)** FAR across the CONUS domain.

Similar to Figure 10, Figure 11 shows spatial verification results, but now for the +2h-QPF model. In contrast to the +1h-QPF configuration, Figure 11a shows F1 decreases more noticeably across portions of the central and eastern United States. While the highest F1 values remain generally aligned with regions of frequent warning activity, the magnitude of forecast skill is reduced relative to both the No-QPF and +1h-QPF models. The FAR distribution for the +2h-QPF model (Fig. 11b) shows a more pronounced increase in false detections across parts of the eastern United States. Substantially higher FAR values were observed in portions of the Mid-Atlantic and northeastern United States compared to the previous model configurations. Some portion of the elevated FAR may also reflect small spatial offsets between predicted probabilities and warning polygons. The following section examines how the results change when spatial tolerance is introduced into the verification framework.

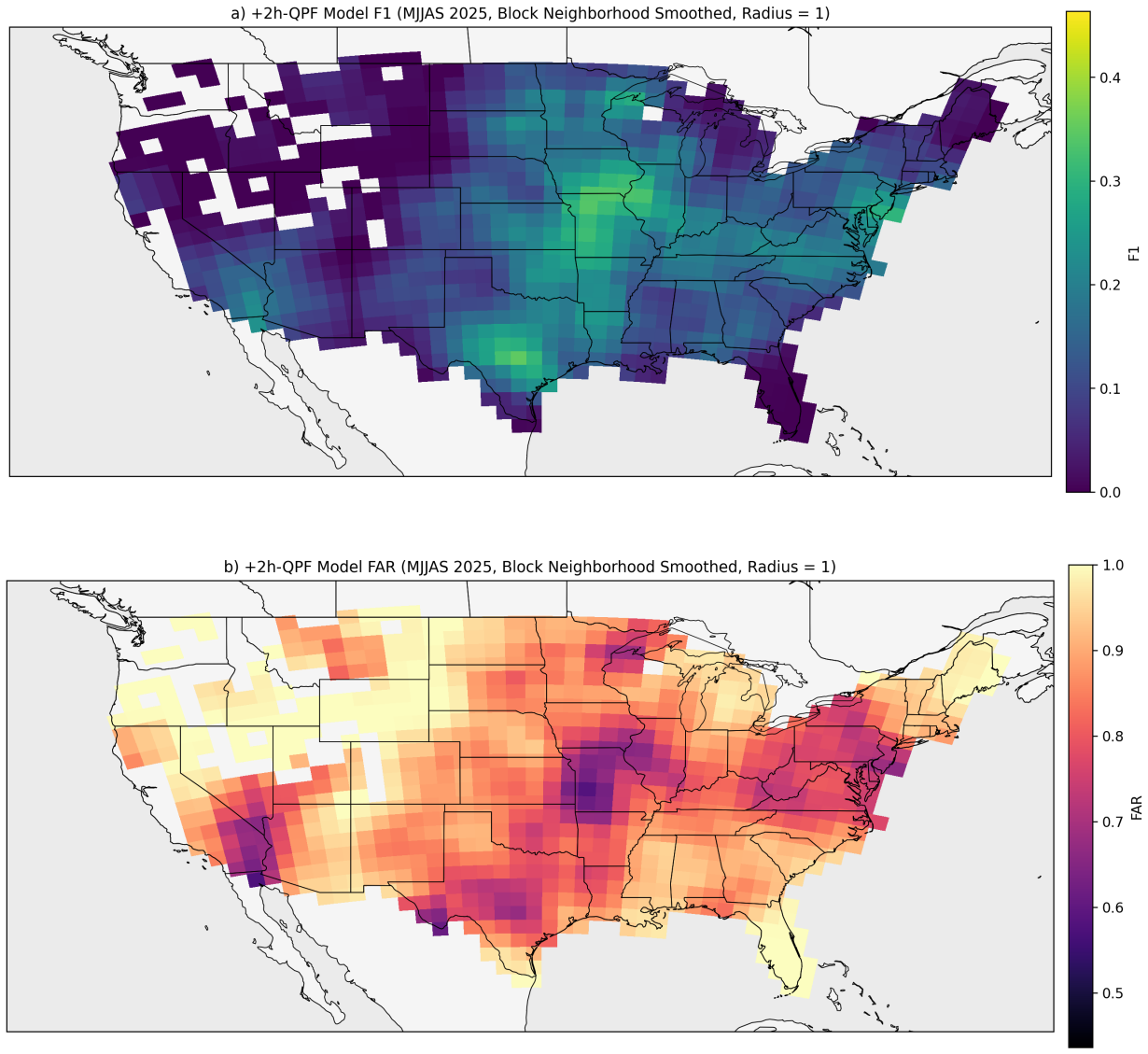

Figure 11. Spatial distribution of forecast skill and false alarm behavior during the MJJAS 2025 testing period for the +2h-QPF model. Verification statistics were calculated within 33x33 grid-cell neighborhoods at the 50% prediction threshold across the CONUS domain and subsequently smoothed using a 1-block radius neighborhood average. **(a)** F1 across the CONUS domain. **(b)** FAR across the CONUS domain.

4.3 Sensitivity to Spatial Verification Tolerance

To examine how strongly the CONUS verification statistics depend on strict pixel-based overlap with warning polygons, additional verification experiments were performed in which the observed warning mask was expanded outward by several grid cells prior to evaluation. This approach was motivated by the observation that forecast probabilities were often centered on warning regions but extended slightly beyond polygon boundaries. Expanding the warning mask provides a simple way to introduce spatial tolerance for these small offsets, similar to neighborhood-based verification methods, but with more direct interpretability. Buffer distances of 2, 4, 6, and 8 pixels were applied to the warning polygons, and FAR was recomputed across forecast probability thresholds for all three model configurations.

Figure 12 shows that FAR decreases systematically as spatial tolerance is relaxed for the No-QPF, +1h-QPF, and +2h-QPF models. This behavior is present across nearly the full range of forecast probability thresholds and is consistent among all three model configurations. The largest reduction in FAR occurs between the zero-buffer control and the first few buffer distances, after which the decrease continues more gradually with additional spatial tolerance. These results indicate that a meaningful fraction of pixels counted as false alarms under the baseline verification framework are only slightly outside of warning boundaries.

This sensitivity to spatial tolerance does not appear to be explained primarily by large spatial displacement errors. In several cases examined (not shown), the model correctly concentrates predicted probabilities near the central region of concern but produces forecasts that appear somewhat broader and more diffused near the edges than the official warning footprint. As a result, the probabilities extend beyond the warning polygon. Under strict pixel-based verification, these peripheral probability fields are counted as false alarms, which inflate FAR

and contribute to the apparent overconfidence of the ML models. Increasing the spatial verification region reduces the apparent false alarms, consistent with edge spillover around otherwise well-placed forecast maxima.

The similarity of this behavior across all three model configurations is important. Although the +1h-QPF and +2h-QPF models maintain higher FAR than the No-QPF model at a given threshold, all three show the same qualitative response to increasing spatial tolerance. This indicates that the influence of strict polygon boundaries on apparent false alarm behavior is not unique to a single model configuration but is instead a broader feature of how these forecasts are being verified against the warning dataset. It is important to note that a by-product of this test is increasing the base rate, so some of the decrease in FAR could simply be due to this factor.

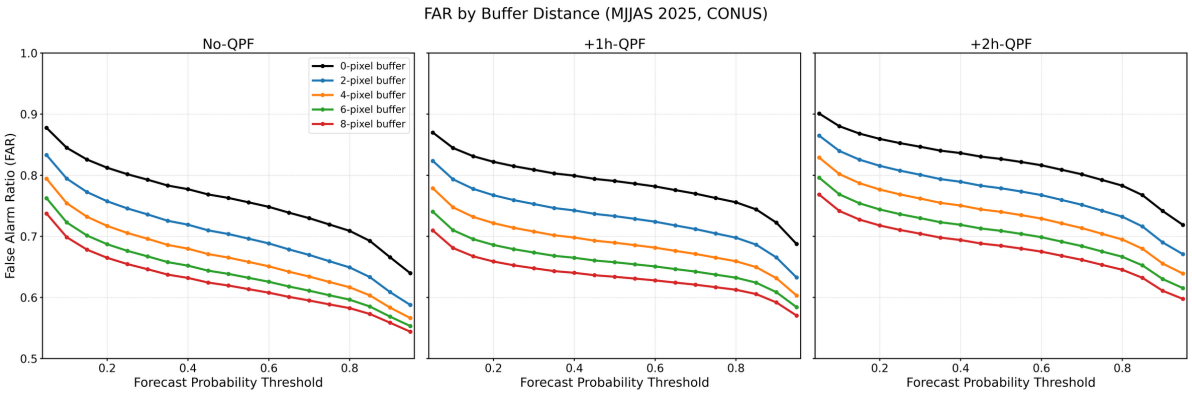

Figure 12. FAR as a function of forecast probability threshold for the No-QPF, +1h-QPF, and +2h-QPF models with increasing spatial verification tolerance during MJJAS 2025 over CONUS. Verification was repeated using warning masks expanded by 2, 4, 6, and 8 pixels.

4.4 NWS Flood Product Verification

Since the NWS communicates flood hazards using products other than FFW, additional verification experiments were performed to examine how the results change when the target definition is broadened beyond FFW alone. In this analysis, contingency statistics were recomputed using a combined target that included FFW, FW, and FA. This alternative dataset was built according to the same methods in Chapter 2.1. This comparison was performed for all three model configurations across the same set of forecast probability thresholds used in the FFW-only verification. While the incorporation of FW and FA products does, to some extent, result in casting an overly broad net, it also includes events for which FW and/or FA products were issued for an event with a flash-flood component (such as events that would be missed when FFW alone are used).

Figure 13 compares verification using FFW alone with verification using the combined FFW, FW, and FA target. For all three model configurations, FAR decreased at every probability threshold when using the new verification dataset. This indicates that the false alarm characteristics of the models depend in part on which NWS flood products are included in the verification definition. It also suggests that some false alarms in the FFW-only framework were associated with alternate operational flood messaging rather than being wholly unsupported forecasts. An event which warrants an FW or FA product might look similar enough to FFW conditions that the models generate non-zero probabilities of FFW issuance. Finally, as with Chapter 4.3, the base rate was inherently increased with this test, which could contribute to lower FAR.

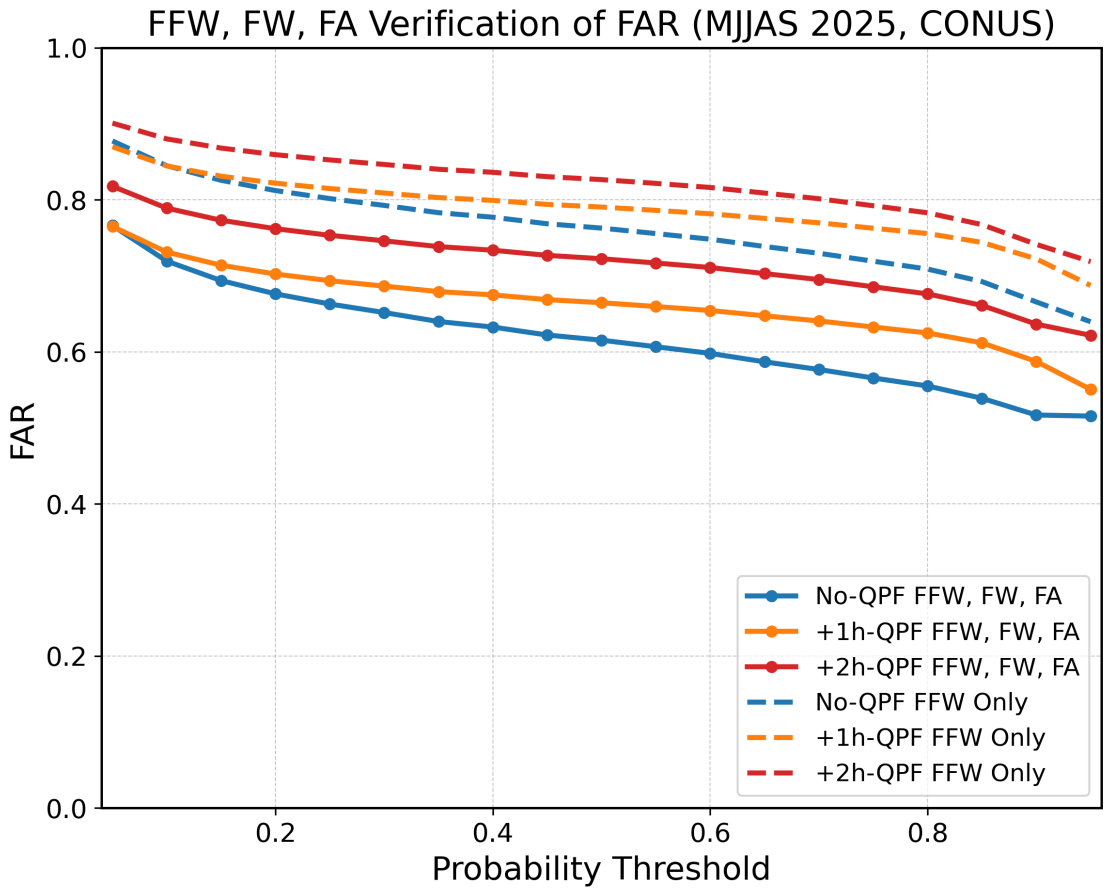

Figure 13. FAR as functions of forecast probability threshold for the No-QPF, +1h-QPF, and +2h-QPF models during MJJAS 2025 over CONUS. Solid lines show verification using a combined target consisting of FFW, FW, and FA, while dashed lines show verification using FFW only.

4.5 Feature Importance

To assess predictor importance, permutation variable importance tests were performed in which one channel or channel group was perturbed during inference while all remaining inputs were left unchanged. To perturb an input field, the original field in each patch was replaced with Gaussian noise randomly generated with the channel specific normalization statistics from

training. This meant the forecast system was rerun with useful information from the perturbed input field effectively removed while preserving the expected magnitude and variability of the perturbed field. Verification statistics were then recomputed from the resulting forecasts to quantify how strongly model performance depended on each predictor group.

Permutation variable importance was applied to groups of physically related channels to improve interpretability and account for strong correlations between related inputs. For example, recent rainfall was evaluated as a combined perturbation of the 0- to 1-hour and 1- to 3-hour accumulation channels, while longer-duration rainfall accumulations were grouped separately to represent antecedent conditions. Many predictors represent the same underlying process distributed across multiple time windows (e.g., rainfall accumulations or QPF), and perturbing a single channel in isolation would allow the model to retain similar information from adjacent channels. Grouping these variables ensures that the importance of the full physical signal is assessed, rather than the marginal contribution of a single, potentially redundant input.

Figure 14 summarizes the threshold dependent response of the No-QPF model to these perturbations. The clearest result is that the model depends strongly on maximum unit streamflow and recent rainfall, specifically the 0- to 1-hour and 1- to 3- hour rainfall channels. Replacing either of these predictor groups with noise produced the largest decreases in F1 score across nearly the full range of probability thresholds. In contrast, perturbations applied to the remaining predictor groups generally produced much smaller reductions, indicating that their contributions to the No-QPF model were significantly lower relative to maximum unit streamflow and recent rainfall.

Figure 14 also shows that the effects of channel removal were not uniform across thresholds. The maximum unit streamflow and recent rainfall varied significantly in their

importance, peaking in importance at the 0.8 threshold. Generally, the other variables became more important with increasing thresholds. This could indicate the model is using all the available information to make its most confident predictions.

A final feature of Figure 14 is that perturbing some predictor groups produced slight increases in F1 score at certain thresholds. Most notably, replacing rainfall from more than 3 hours prior to analysis time with noise led to modest F1 increases across the lower half of the threshold range, meaning this variable is detracting from model performance.

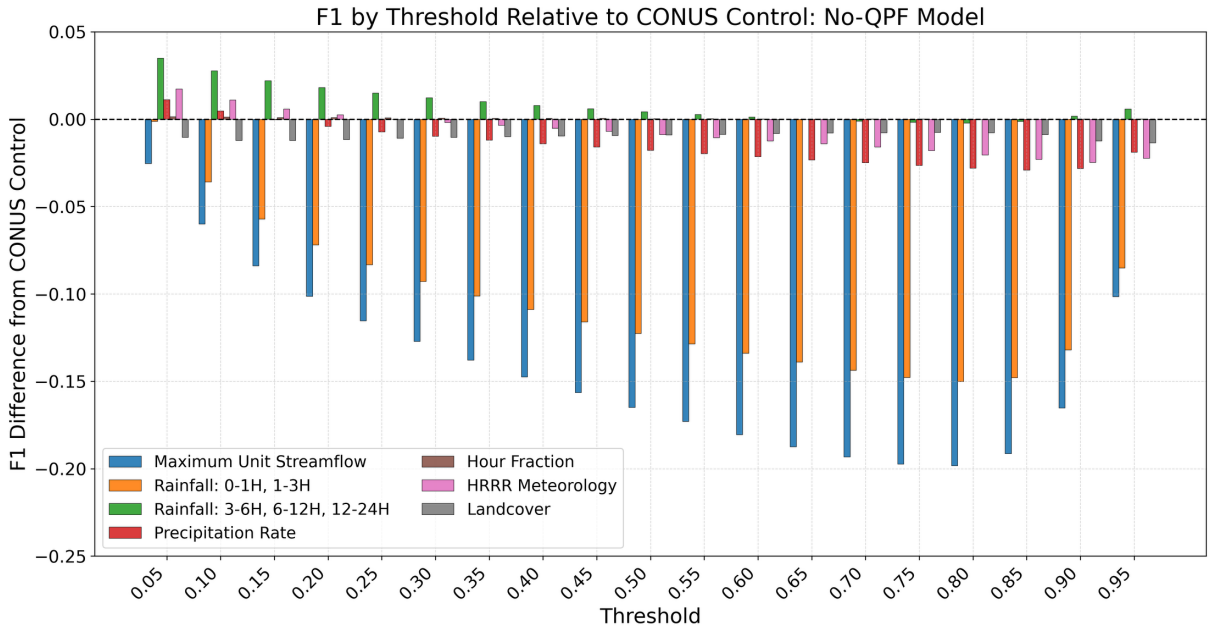

Figure 14. Difference in F1 score from the CONUS control as a function of forecast probability threshold for the No-QPF model in the permutation variable importance tests. Each colored bar represents the change in F1 score after replacing an individual channel or channel group with Gaussian noise while leaving all other inputs unchanged. Negative values indicate degraded performance relative to the control, while positive values indicate slight improvement.

Figure 15 shows the same threshold-dependent F1 analysis for the +1h-QPF and +2h-QPF models. Generally, the addition of QPF does not substantially alter the broader importance structure from the No-QPF model. In both Figure 15a and 15b, maximum unit streamflow and recent rainfall remain the dominant predictors, while replacing the remaining channel groups with noise produced comparatively smaller impacts. Thus, the inclusion of QPF does not change the primary result from Figure 14.

The QPF perturbations themselves show a threshold-dependent effect. In both the +1h-QPF and +2h-QPF models, replacing the QPF inputs with noise leads to slight F1 increases at lower probability thresholds and F1 decreases at upper thresholds. This indicates that QPF contributes differently across the forecast probability range.

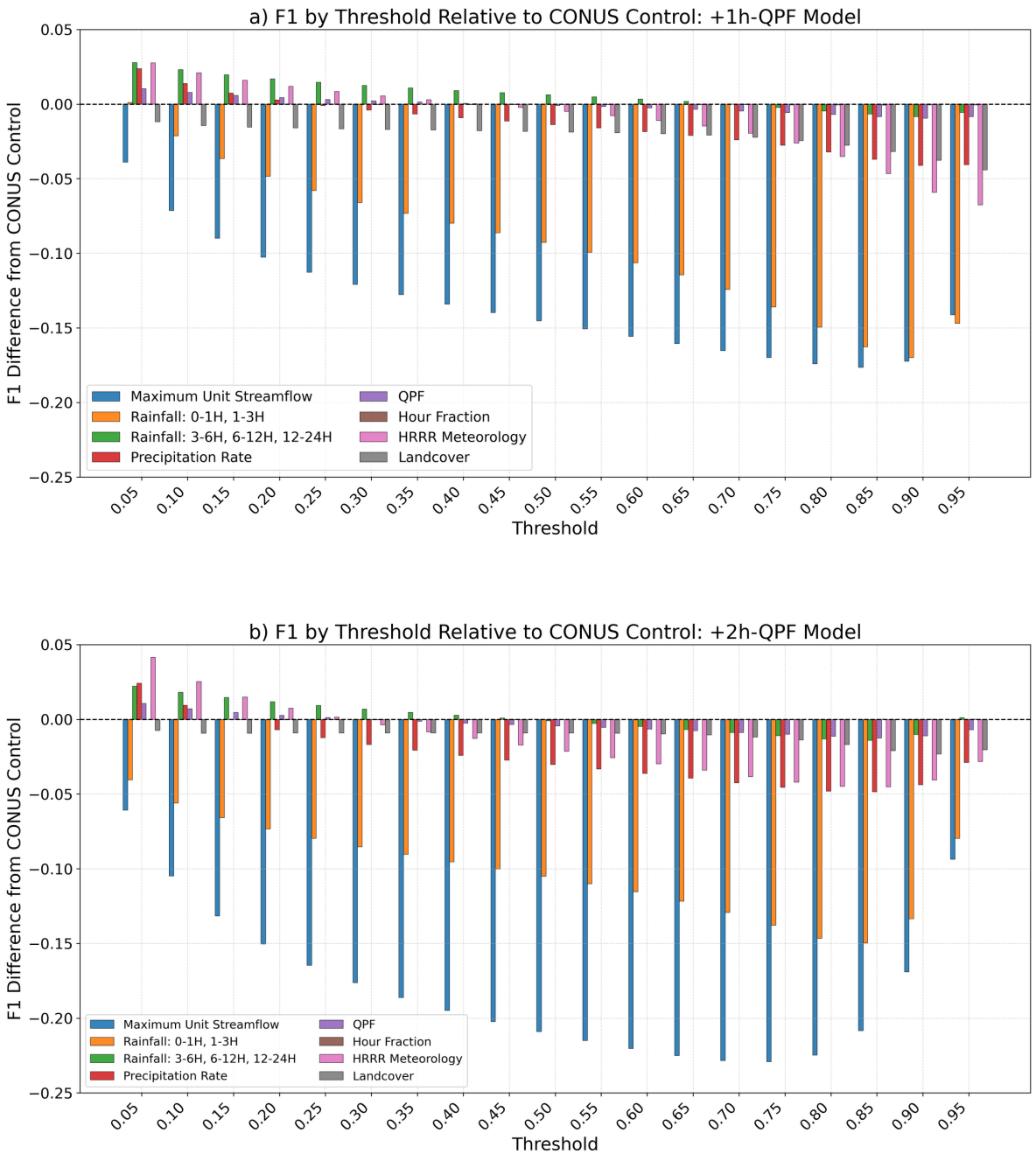

Figure 15. Difference in F1 score from the CONUS control as a function of forecast probability threshold for the permutation variable importance tests applied to the QPF-based models. Each colored bar represents the change in F1 score after replacing one channel or channel group with Gaussian noise while leaving all other inputs unchanged. Negative values indicate degraded

performance relative to the control. **(a)** Shows results for the +1h-QPF model, where noise is applied to the 1-hour QPF channel. **(b)** Shows results for the +2h-QPF model, where noise is applied to the 1- and 2-hour QPF channels.

Table 1 provides an aggregate summary of the permutation variable importance tests using PR-AUC. Unlike the bar charts in Figures 14 and 15, PR-AUC condenses performance into a single summary measure over the full precision-recall curve. As a result, the table confirms which predictor groups matter most in an overall sense. Consistent with the threshold dependent F1 results, the largest reductions in PR-AUC occurred when maximum unit streamflow and recent rainfall were replaced with noise. This behavior was evident across all three model configurations, confirming that these predictors provide the strongest overall contribution to model skill. Conversely, perturbing landcover, hour fraction, HRRR meteorology, and longer duration rainfall generally produced much smaller PR-AUC changes, indicating that these inputs had a more limited influence on the models' predictions. For the QPF-based models, replacing the QPF channels with noise produced modest reductions in PR-AUC, suggesting that forecast precipitation contributed to aggregate skill but certainly less than streamflow and recent rainfall.

Noise Group	No-QPF PR-AUC	No-QPF % Difference	+1h-QPF PR-AUC	+1h-QPF %	+2h-QPF PR-AUC	+2h-QPF % Difference
Maximum Unit Streamflow	0.058646	-57.213%	0.051975	-58.750%	0.040702	-72.715%
Rainfall: 0-1H, 1-3H	0.060278	-56.023%	0.084488	-32.946%	0.080138	-46.277%
Rainfall: 3-6H, 6-12H, 12-24H	0.129122	-5.795%	0.120755	-4.163%	0.140200	-6.014%
Precipitation Rate	0.120619	-11.999%	0.112950	-10.357%	0.119200	-20.091%
Hour Fraction	0.137178	0.083%	0.125511	-0.388%	0.148866	-0.204%
HRRR Meteorology	0.132200	-3.550%	0.130995	3.964%	0.131370	-11.933%
Landcover	0.132380	-3.418%	0.115683	-8.188%	0.142901	-4.203%
QPF			0.123486	-1.995%	0.144164	-3.356%

Table 1. PR-AUC for the permutation variable importance tests applied to the No-QPF, +1h-QPF, and +2h-QPF models during MJJAS 2025 over CONUS. Each row corresponds to a channel or channel group that was replaced with Gaussian noise while all other inputs were retained. Percent differences are reported relative to the corresponding CONUS control for each model. Larger negative percent differences indicate greater sensitivity of model performance to that predictor group.

Chapter 5: Test Cases

To complement the aggregate verification statistics, two warm season case studies were examined to assess how the models behave during events of different regimes. The first case, 4 July 2025, represents a widespread, high impact flash flood event with numerous flood related warning products and a large spatial footprint. The second case, 25 to 26 May 2025 is a more modest and spatially fragmented event. Together, these cases illustrate four recurring features of the model guidance:

1) The ability to identify the primary corridor of flood risk
2) Earlier signal development in the QPF-based configurations
3) Better alignment of high probabilities with FFW, FW, and FA than with FFW alone
4) Larger spatial footprint of model predictions than the polygons themselves

5.1 4 July 2025

On 4 July 2025, a prolonged heavy rainfall event produced significant flash flooding across central Texas, with the most severe impacts concentrated in Kerr County and surrounding

areas. The event evolved during the overnight hours as a slow-moving convective complex developed and remained over the same region for an extended period. Initial flood-related products were issued on the night of July 3rd, with FFWs issued shortly thereafter as rainfall intensified and impacts became more apparent. The most significant flooding began around 0800 UTC, when convection over Kerr County led to rapid runoff into the Guadalupe River creating life threatening conditions. As the event progressed, a combination of FFW, FW, and FA were used to communicate risk.

By 0430 UTC (Fig. 16a-c) on 4 July 2025, all three models had already identified the initial area of concern in central Texas. In each configuration, forecast probabilities approached 100% in the vicinity of the first FFW. Two FW/FA were also present in the southern portion of the panel. In that area, the models indicated a 70-90% chance of a FFW, but the local forecast office chose an alternative flood product. Notably, forecast confidence in this southern area increased as QPF information was added, suggesting that the inclusion of future precipitation imparted improved ML model predictions.

At 0600 UTC (Fig. 16e-g), shortly before the first FFW in Kerr County was issued, the +1h-QPF and +2h-QPF models were more confident in warning-level conditions than the No-QPF model. It is also notable that, in the northwestern warning region, the +2h-QPF model did not expand broad areas of lower probability to the same extent as the No-QPF and +1h-QPF models. This suggests that the stronger signal in the +2h-QPF model was not simply the result of uniformly higher forecast probabilities but instead reflected more targeted forecasts in areas that were later associated with warning activity.

At 0800 UTC (Fig. 16i-k), all three models predict FFW with very high confidence over Kerr County, where the most devastating flooding occurred. A second local maximum in

probabilities was also evident on the northwestern flank of the convective complex. This region was later associated with a flash flood emergency. The forecasts therefore showed skill not only in identifying the broader flood corridor, but also in reserving the highest probabilities for the most severe warning regimes.

Finally, by 1030 UTC (Fig. 16m-o), the convective complex had expanded to cover a much larger area, with the heaviest precipitation on the western edge. The models maintained the highest probabilities along the western side of the complex, consistent with the area of greatest ongoing concern. At this time, the FFW over Kerr County was no longer in effect. Instead, the forecast office relied primarily on FW and FA to cover much of the area where the model continued to indicate high FFW probabilities. Within the following two hours, a new FFW was issued in western Kerr County, as represented by the blue polygon in Figure 16o. Thus, although FFW were the primary product during the event, other flood-related products created a patchwork of flood risk that cannot be captured using FFW alone.

As the event evolved, the forecasts from all three models expanded into a broad and spatially coherent swath that overlapped much of the observed flood impact corridor. This consistency across model configurations is important because it indicates that the guidance was not simply responding to isolated convective pixels but was capturing the larger scale organization of the event. The +1h-QPF and +2h-QPF models generally developed this broader signal earlier than the No-QPF model, with the +2h-QPF configuration showing the strongest downstream extension. In an operational setting, this earlier development could provide additional situational awareness before the full warning footprint becomes apparent from observations alone.

This case also highlights the limitations from Chapter 4. First, notice how nearly all of the warning probabilities from the model are maximized over FFW polygons, but also spill over into regions outside of the polygons. Second, in several panels, high forecast probabilities were collocated with other flood-related warning products rather than with FFW alone. Under the baseline verification framework in Chapter 4.1, these areas would be counted as false alarms, even though the forecast signal still corresponded to meaningful flood-related impacts and operational response.

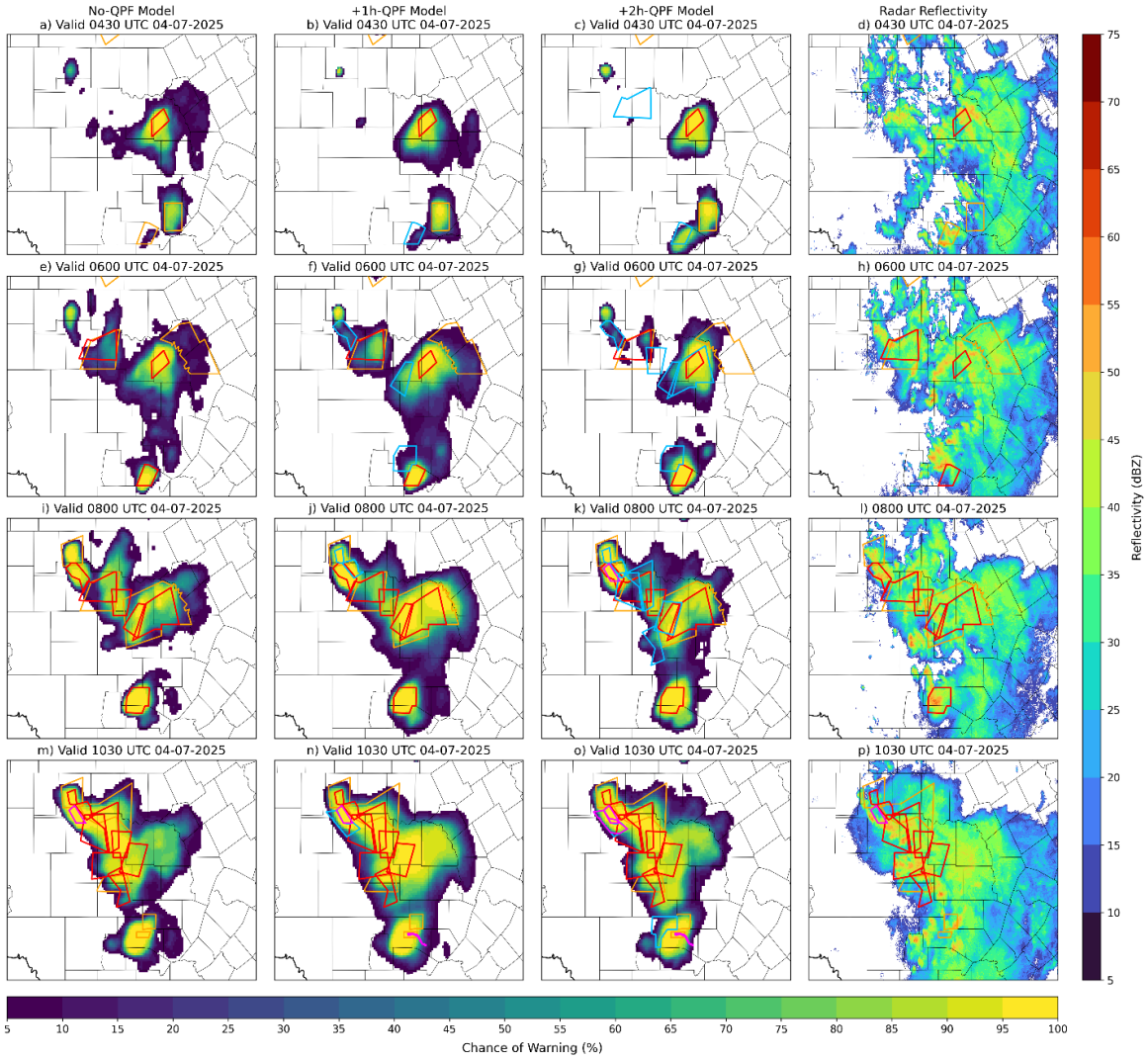

Figure 16. Model output from **a-c)** 0430, **e-g)** 0600, **i-k)** 0800, and **m-o)** 1030 UTC on 4 July 2025, in south-central Texas. NWS polygons plotted include flash flood emergencies (fuchsia), flash flood warnings (red), future flash flood warnings within the lead times considered by each model (blue), and flood warning or flood advisory (orange). **d, h, i, p)** Radar reflectivity at analysis time.

5.2 25-26 May 2025

The second case occurred from 25 to 26 May 2025 in south central Oklahoma. This event provides a useful contrast to the 4 July case because it was more localized and spatially fragmented. Despite its smaller scale, all three models still identified the principal areas of concern. By 2200 UTC (Fig. 17a-c), the +2h-QPF model had already highlighted a region in the Oklahoma City metropolitan area that was later encompassed by FFWs. Relative to the +1h-QPF model, the +2h-QPF configuration showed a clearer northward and eastward extension of probabilities, which was more consistent with the warning evolution in Figures 17e-g. In contrast, the No-QPF model showed little to no signal over the Oklahoma City area at this time. This implies, to an extent, that the +2h-QPF and +1h-QPF models were inferring about the future rather than just the current state. It is also notable that an alternative flood product was used in southern Oklahoma where the models indicated a high likelihood of FFW conditions. This could mean that either the conditions the ML models have learned correspond to other types of flooding, or that flash flooding was possible when other types of products were issued.

By 2300 UTC (Fig. 17e-g), the first FFW in the Oklahoma City metropolitan area had been issued, and all three models captured that area of concern. At this stage, one of the main differences among the configurations was the southward expansion of forecasted risk. The +2h-QPF model showed higher probabilities to the south than the other models suggesting an anticipation of the evolving flood threat, which does ultimately verify according to the blue polygon south of the active warning polygon in Figure 17g. The +1h-QPF model also maintained a southern area of elevated probabilities more persistently than the other models, which could be a false alarm or a tendency of the model to retain probabilities too long after events have ended.

At 0300 UTC of 26 May (Fig. 17i-k), the flash flood threat in southern Oklahoma began to increase. Once again, the +1h-QPF and +2h-QPF models showed higher probabilities than the No-QPF model in areas that were about to be warned, suggesting that the QPF information used by the +1h-QPF and +2h-QPF versions of the model positively impacts the models' ability to identify future areas of FFW issuance. Of particular interest, the +2h-QPF model showed a more pronounced southeastward extension of probabilities, consistent with the later propagation of the convective complex toward the Red River region. This southeastward signal was stronger than in the +1h-QPF model, suggesting that the additional forecast precipitation contributed meaningful lead time and was not simply increasing probabilities uniformly across the domain.

By 0350 UTC (Fig. 17m-o), the No-QPF model had also developed a clear signal where a FFW was issued near the Red River. As in the previous panel, each configuration with added forecast information progressively highlighted areas farther to the southeast as having elevated risk of FFW issuance. In addition, all three models showed very high confidence in the warned area itself, generating regions of high probability of FFW issuance that closely resembled the issued warning polygons. This indicates that the models maintained strong spatial precision even in a relatively compact flood event. At the same time, the placement of other flood-related products did not always align with forecast probabilities, which complicates the interpretation somewhat and suggests that additional work is needed to better understand how flood and flash flood hazards are represented in ML training and validation datasets.

Overall, the 25 to 26 May 2025 case demonstrates that the models retained useful skill even in a smaller, more fragmented flood event. All three configurations identified the main areas of concern, but the +1h-QPF and +2h-QPF models generally developed stronger and earlier signal in regions that were later covered by FFW. The +2h-QPF model was especially effective at

identifying future regions of concern, indicating that the added precipitation forecast information contributed meaningful lead time rather than simply increasing probabilities everywhere. At the same time, this case again illustrates two important themes that appear throughout the study. First, some (but not all) high probability forecast areas corresponded to other flood-related products rather than FFW alone. Second, in a spatially compact event, the tendency of the ML models to predict regions of probability that extend outside polygon boundaries will strongly impact pixel-based verification.

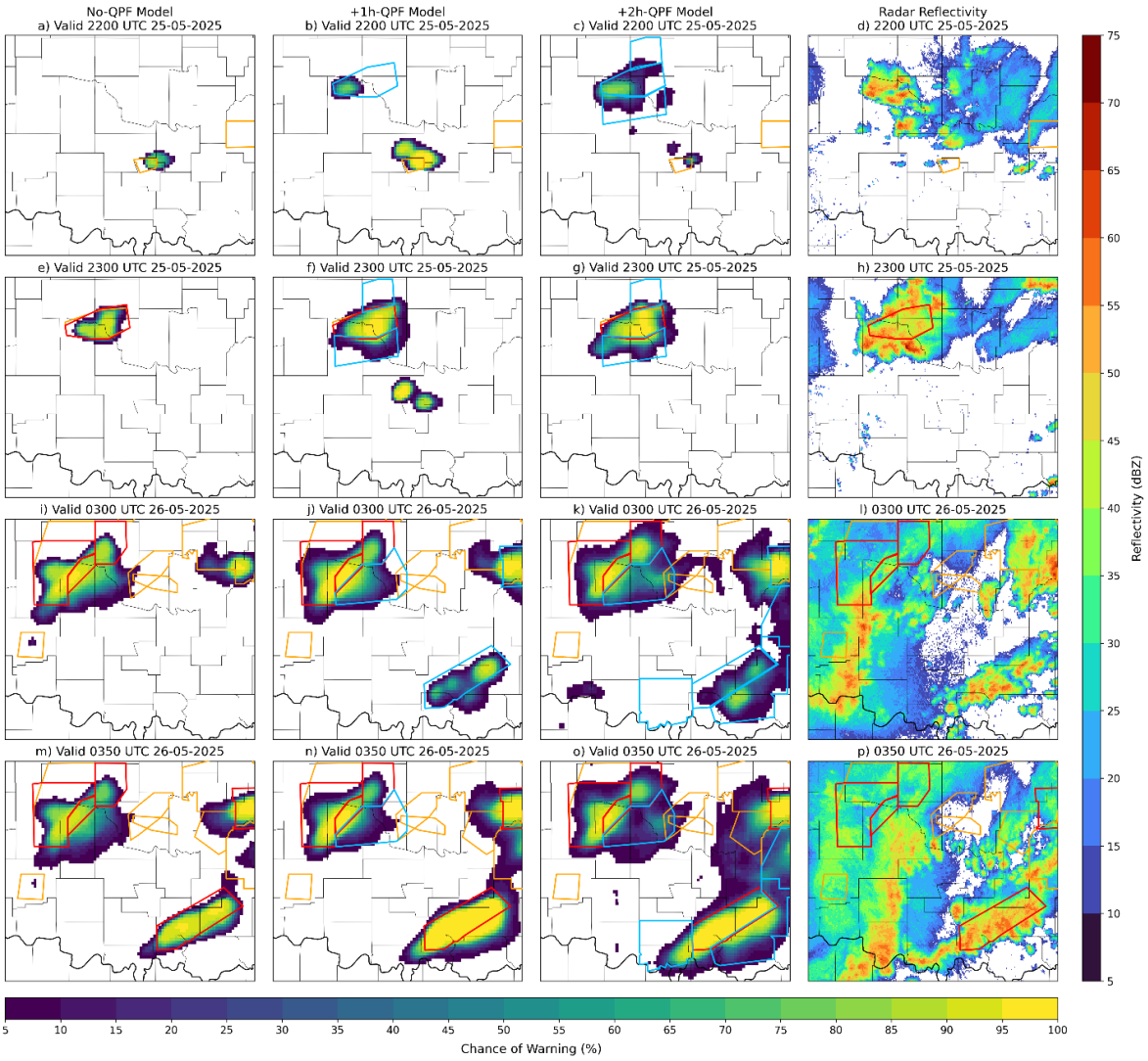

Figure 17. Model output from **a-c)** 2200 UTC and **e-g)** 2300 UTC on 25 May 2025 and **i-k)** 0300 UTC and **m-o)** 0350 UTC on 26 May 2025 over south-central Oklahoma. NWS polygons plotted include flash flood emergencies (fuchsia), flash flood warnings (red), future flash flood warnings within the lead times considered by each model (blue), and flood warning or flood advisory (orange). **d, h, i, p)** Radar reflectivity at analysis time.

Chapter 6: Discussion

6.1 Interpretability

Permutation variable importance tests provide a useful perspective on what information the models relied on most heavily when producing FFW forecasts, allowing us to "look inside the ML black box". Across all three model configurations, the clearest result was the dominant importance of maximum unit streamflow and recent rainfall. Using noise to deny either of these predictor groups produced the largest reductions in skill, both in the threshold dependent F1 analysis and in the aggregate PR-AUC results. This is a physically reasonable outcome, since FFW issuance is strongly tied to the combination of current hydrologic state and ongoing or very recent rainfall. Therefore, the models appear to be learning a signal that is consistent with the underlying processes that govern FFW issuance.

Part of this strong dependence on maximum unit streamflow is also physically consistent with the construction of the product itself, since it already reflects the influence of terrain, land use, and hydrologic response characteristics. Some of the information contained in the separately provided land use predictors may already be represented, to a degree, within the streamflow input, making the models less reliant on those predictors individually. This is suggested by the permutation importance results where land use was not relied on heavily by the model.

Examining threshold-dependent behavior provides a more nuanced view of how the models use the information provided to them. For maximum unit streamflow and recent rainfall, the largest performance reductions occurred at middle to upper probability thresholds, with the strongest impacts generally occurring near a threshold of 0.80. This suggests that these predictors are especially important for the model's high confidence forecasts. By contrast, several other

predictor groups produced smaller and more threshold dependent responses, indicating that their contributions were context dependent.

The importance tests also suggest that not all useful predictors contribute in the same way. In some cases, perturbing longer duration antecedent rainfall produced slight F1 increases at lower thresholds. One possible explanation is that the models coincidentally learn an association between longer-duration antecedent rainfall and FFW. While this antecedent rainfall can contribute to a more favorable hydrologic background state, it is not necessarily the direct signal driving warning issuance at analysis time. Consequently, the model may use these inputs to paint broader areas of low probability tied to saturated soils rather than more focused regions of imminent FFW risk. This behavior is consistent with the elevated FAR and reduced precision observed at lower probability thresholds.

A similar conclusion applies to the QPF-enabled models. Including QPF did not substantially alter the overall ranking of predictor importance, as maximum unit streamflow and recent rainfall remained the dominant contributors in both the +1h-QPF and +2h-QPF configurations. However, the threshold dependent F1 results suggest that QPF was not applied equally within the model. At lower thresholds, QPF may contribute a broad background signal associated with general convective potential, slightly degrading threshold-specific F1. With stronger signals in QPF, the models seem to rely more heavily on this information to make skillful high-probability predictions.

6.2 Regional, Spatial, and Verification Target Limitations

One of the clearest themes in the CONUS verification results is a substantial over-forecasting bias when performing strict pixelwise FFW-only evaluation across the full domain. The regional verification results show that the apparent over-forecasting bias varies considerably across the domain, with some regions contributing disproportionately to the national FAR. Florida, and to a lesser extent parts of the Northwest, exhibit much higher FAR than most of the central and eastern United States. This could be due to the high observed warning count in the central and eastern United States, which means the model's training data likely included many cases from these regions. This would result in models that were trained to identify patterns common to that part of CONUS which are not necessarily common in other areas.

The spatial tolerance tests further suggest that part of this high FAR is tied to the spatial structure of the forecasts rather than solely to misplaced predictions. Allowing a buffer around warning polygons reduced FAR across all three model configurations, indicating that some forecast pixels classified as false alarms under strict verification were located close to the edges of warned regions. This supports the interpretation that the models often identify the central region of concern but produce somewhat broader and more diffuse probability footprints around the polygon boundaries. However, the impact of base rate needs to be further investigated to understand if this, and the next point, are conclusive findings.

Most importantly, the verification tests using the combined FFW, FW, and FA target suggest that the issue is not limited to spatial spillover alone. When FW and FA were included in the verification target, FAR decreased systematically for all three models. While this does not indicate a direct prediction of these products, it highlights a deeper issue with this research topic. These models and forecast verification are using one specific product, even though forecasters

have several products available to communicate flooding risk. Defining a label dataset that accurately and completely captures forecasters' perspectives on where flash flooding was occurring or likely to occur is a difficult challenge, and one where future work vital for ML studies remains needed.

This is the central target problem in the study. FFWs are a useful operational label, but they are not a complete spatial and temporal representation of flood risk. They reflect a human decision made within a warning framework that includes multiple products, local office practices, and event-specific judgement. A more appropriate target for this problem may be something closer to a cohesive flood-analysis field, analogous to the SPC's practically perfect or the NHC's best track frameworks. The goal of these concepts is to reconstruct the most realistic spatial and temporal depiction of the event after it has occurred. The results presented here suggest that the models may already be learning a broader flood signal. The challenge is that the current training and verification target does not fully represent it.

Another approach that could be taken is to use FLASH CREST maximum unit streamflow as the target variable using the other predictors from this study and take out FFWs. By doing this, the human factor is eliminated, and prediction becomes much more objective. From this, forecasters could then deduce flood product issuance based on this product, instead of trying to predict what the forecaster will do. Predicting maximum unit streamflow presents its own challenges, though, as it uses a 1km grid (at the spatial scale of fine/noisy processes) and varies significantly from grid point to grid point.

6.3 Operational Use

Although the present system is not sufficiently mature for operational deployment, the results demonstrate a promising approach for filling the watch-to-warning timescale in flash flood forecasting. The models produce probabilistic guidance that may offer useful early indication of high-probability flood threats before FFW issuance. The most realistic near-term role of this system is as an experimental decision support tool rather than operational guidance.

It is important to consider that the results suggest operational value should not be judged solely by exact overlap with FFW polygons. The model often identifies coherent regions of elevated flood concern (especially in the central and eastern United States), but these probability footprints can be broader than the final warning area and may also align with alternate flood-related products.

Substantial development would still be required before such a system could support real-time operations. Important next steps could include regional training and validation, improved handling of false alarm behavior, refinement of the target definition, and testing within a real-time forecasting environment. Forecaster-centered evaluation is essential since the practical utility of the guidance depends on whether it improves awareness, confidence, and lead time in real warning decision settings.

Chapter 7: Conclusion

This thesis examined the potential for machine learning to support short-term flash flood forecasting by predicting the probability of NWS FFW issuance at 0- to 2-hour lead times. Using a U-Net framework trained on MRMS rainfall, FLASH maximum unit streamflow, HRRR

meteorological fields, land cover, and optional HRRR QPF, the models demonstrated meaningful skill during the 2025 warm season test period across the CONUS domain. Among the three configurations, the No-QPF model produced the strongest overall precision-recall performance, while the QPF-based models generally showed lower precision and higher FAR. However, the inclusion of forecast precipitation provided earlier signal development in several situations, indicating a tradeoff between aggregate skill and potential lead time.

A central finding of this work is that the models learned a physically coherent signal. Permutation variable importance tests showed that maximum unit streamflow and recent rainfall were the dominant contributors to skill across all model configurations, while other predictor groups had smaller and more threshold-dependent effects. This result is encouraging because it suggests that the models are relying most heavily on information directly tied to the hydrologic and rainfall conditions that govern flash flooding. QPF played a more nuanced role, contributing less to aggregate PR-AUC than streamflow and recent rainfall, but appearing more useful at middle to upper probability thresholds.

The verification results showed that apparent over-forecasting must be interpreted carefully. Strict FFW-only verification across CONUS produced substantial FAR, but that signal was not spatially uniform and was disproportionately influenced by regions such as Florida and parts of the Northwest. Sensitivity to spatial tolerance tests further showed that a higher tolerance meant reduced FAR, indicating that some false alarms could be tied to probabilities extending beyond warning edges. Verification using a broader target that included FW and FA also produced more favorable statistics than verification against FFW alone. This suggests that part of what is labeled false alarm under strict FFW-only verification may reflect a broader flooding signal that forecasters communicate using multiple operational products. More broadly, the

results point to a target-definition problem, where FFW polygons are useful operational labels, but they may be too narrow to fully represent the spatial and temporal structure of flood risk.

Overall, this study demonstrates that machine learning offers a promising pathway for bridging the watch-to-warning timescale in flash flooding. The results show that ML can identify meaningful areas of elevated flood concern and can provide probabilistic guidance tied to physically relevant hydrologic and rainfall signals. Although the system is not ready for operations, it shows clear potential as a decision support tool that could help forecasters in the future.

Several important areas for future work remain. Considering the strong regional dependence, individual models could be tested and verified in regions of the United States rather than all of CONUS. This would perhaps reduce FAR locally. Further work could also examine more spatially and temporally cohesive target definitions, potentially moving beyond strict FFW polygons toward a broader representation of flood risk. Expanded evaluation in a real time forecast environment would also be important to understanding forecaster reception. Changing the class balance of training patches or calibrating the final data could help reduce some of the remaining systematic overconfidence. Finally, it would be valuable to do further investigation into the role of base rates in verification statistics. These avenues of exploration show the breadth and complexity of this project, and that its continued development extends well beyond one analysis.

References

American Meteorological Society, 2026: *Glossary of Meteorology*, accessed 13 January 2026, https://glossary.ametsoc.org/wiki/Flash_flood.

Archer, D. R., and H. J. Fowler, 2018: Characterising flash flood response to intense rainfall and impacts using historical information and gauged data in Britain. *J. Flood Risk Manag.*, **11**, https://doi.org/10.1111/jfr3.12187.

Chase, R. J., D. R. Harrison, G. M. Lackmann, and A. McGovern, 2023: A Machine Learning Tutorial for Operational Meteorology. Part II: Neural Networks and Deep Learning. *Weather Forecast.*, **38**, 1271–1293, https://doi.org/10.1175/WAF-D-22-0187.1.

Clark, A. J., and E. D. Loken, 2022: Machine Learning–Derived Severe Weather Probabilities from a Warn-on-Forecast System. *Weather Forecast.*, **37**, 1721–1740, https://doi.org/10.1175/WAF-D-22-0056.1.

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products. *Weather Forecast.*, **29**, 377–392, https://doi.org/10.1175/WAF-D-12-00124.1.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash Flood Forecasting: An Ingredients-Based Methodology. *Weather Forecast.*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011%253C0560:FFFAIB%253E2.0.CO;2.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. *Weather Forecast.*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Fingerhut, H., 2025: Timeline raises questions over how Texas officials handled warnings before the deadly July 4 flood. *Associated Press*, July 11.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using Machine Learning to Generate Storm-Scale Probabilistic Guidance of Severe Weather Hazards in the Warn-on-Forecast System. *Mon. Weather Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

——, P. Skinner, C. K. Potvin, B. Matilla, and A. Reinhart, 2025: Assessing the Impact of Biased Target Variables on Machine Learning Models of Severe Hail. *Weather Forecast.*, **40**, 1015–1028, https://doi.org/10.1175/WAF-D-24-0051.1.

Gerard, A., S. M. Martinaitis, J. J. Gourley, K. W. Howard, and J. Zhang, 2021: An Overview of the Performance and Operational Applications of the MRMS and FLASH Systems in Recent Significant Urban Flash Flood Events. *Bull. Am. Meteorol. Soc.*, **102**, E2165–E2176, https://doi.org/10.1175/BAMS-D-19-0273.1.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Weather Forecast.*, https://doi.org/10.1175/WAF-D-21-0026.1.

——, ——, and M. L. Green, 2024: Observation Definitions and Their Implications in Machine Learning–Based Predictions of Excessive Rainfall. *Weather Forecast.*, **39**, 1733–1750, https://doi.org/10.1175/WAF-D-24-0033.1.

Li, J., L. Li, T. Zhang, H. Xing, Y. Shi, Z. Li, C. Wang, and J. Liu, 2024: Flood forecasting based on radar precipitation nowcasting using U-net and its improved models. *J. Hydrol.*, **632**, 130871, https://doi.org/10.1016/j.jhydrol.2024.130871.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, 2017: Hyperband: Bandit-Based Configuration Evaluation For Hyperparameter Optimization. International Conference on Learning Representations.

Martinaitis, S. M., and Coauthors, 2023: A Path Toward Short-Term Probabilistic Flash Flood Prediction. *Bull. Am. Meteorol. Soc.*, **104**, E585–E605, https://doi.org/10.1175/BAMS-D-22-0026.1.

Mosavi, A., P. Ozturk, and K. Chau, 2018: Flood Prediction Using Machine Learning Models: Literature Review. *Water*, **10**, 1536, https://doi.org/10.3390/w10111536.

National Weather Service, 2024: Weather Related Fatality and Injury Statistics, accessed 27 January 2026, https://www.weather.gov/hazstat.

——, 2025: Flood Watch, issued 1818 UTC 3 July 2025 by NWS Austin/San Antonio (EWX), https://mesonet.agron.iastate.edu/p.php?pid=202507031818-KEWX-WGUS64-FFAEWX.

——, 2026: *National Weather Service Glossary*, accessed 2 February 2026, https://forecast.weather.gov/glossary.php?

Nevo, S., and Coauthors, 2022: Flood forecasting with machine learning models in an operational framework. *Hydrol. Earth Syst. Sci.*, **26**, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022.

NOAA National Centers for Environmental Information, 2024: U.S. Billion-Dollar Weather and Climate Disasters. NOAA National Centers for Environmental Information, accessed 13 January 2026, https://doi.org/10.25921/STKW-7W73.

——, 2026: Storm Events Database, accessed 27 January 2026, https://www.ncei.noaa.gov/stormevents/.

Pielawski, N., and C. Wählby, 2020: Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLOS ONE*, **15**, e0229839, https://doi.org/10.1371/journal.pone.0229839.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/ARXIV.1505.04597.

Schumacher, R. S., A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From Random Forests to Flood Forecasts: A Research to Operations Success Story. *Bull. Am. Meteorol. Soc.*, **102**, E1742–E1755, https://doi.org/10.1175/BAMS-D-20-0186.1.

Seo, D., T. Lakhankar, J. Mejia, B. Cosgrove, and R. Khanbilvardi, 2013: Evaluation of Operational National Weather Service Gridded Flash Flood Guidance over the Arkansas Red River Basin. *JAWRA J. Am. Water Resour. Assoc.*, **49**, 1296–1307, https://doi.org/10.1111/jawr.12087.

Soares, J. A. J. P., L. C. S. M. Ozelim, L. Bacelar, D. B. Ribeiro, S. Stephany, and L. B. L. Santos, 2025: ML4FF: A machine-learning framework for flash flood forecasting applied to a Brazilian watershed. *J. Hydrol.*, **652**, 132674, https://doi.org/10.1016/j.jhydrol.2025.132674.

Stensrud, D. J., and Coauthors, 2009: Convective-Scale Warn-on-Forecast System: A Vision for 2020. *Bull. Am. Meteorol. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmospheric Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Terti, G., I. Ruin, S. Anquetin, and J. J. Gourley, 2017: A Situation-Based Analysis of Flash Flood Fatalities in the United States. *Bull. Am. Meteorol. Soc.*, **98**, 333–345, https://doi.org/10.1175/BAMS-D-15-00276.1.

United States Geological Survey, 2025: Annual National Land Cover Database (NLCD) Collection 1 Products (ver. 1.1, June 2025). U.S. Geological Survey, accessed 4 November 2025, https://doi.org/10.5066/P94UXNTS.

Weather Prediction Center, 2026: WPC Excessive Rainfall Forecast Archive, accessed 13 January 2026, https://www.wpc.ncep.noaa.gov/archives/web_pages/ero/ero.shtml.

Xu, H., H. He, Y. Zhang, L. Ma, and J. Li, 2023: A comparative study of loss functions for road segmentation in remotely sensed road datasets. *Int. J. Appl. Earth Obs. Geoinformation*, **116**, 103159, https://doi.org/10.1016/j.jag.2022.103159.

Yingkai (Kyle) Sha, 2022: yingkaisha/keras-unet-collection: v0.1.13, https://doi.org/10.5281/ZENODO.5834880.

Yussouf, N., K. A. Wilson, S. M. Martinaitis, H. Vergara, P. L. Heinselman, and J. J. Gourley, 2020: The Coupling of NSSL Warn-on-Forecast and FLASH Systems for Probabilistic Flash Flood Prediction. *J. Hydrometeorol.*, **21**, 123–141, https://doi.org/10.1175/JHM-D-19-0131.1.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Am. Meteorol. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.