Improving EFAS's Decision-Making on Fluvial Flood Warning Notifications with Machine Learning

MSc Thesis

(Bianchi, 2021)

Name:	Noemi Barbieri
Registration number:	1356992
Main Supervisor:	Samuel Sutanto
Second Supervisor:	Shahab Torbaghan
Study Programme:	MSc Urban Environmental Management
Chair:	Water System and Climate Change

WAGENINGEN
UNIVERSITY & RESEARCH

Abstract

River floods are among the most destructive natural hazards globally, making timely and accurate Early Warning Systems essential for disaster risk reduction. European Flood Awareness Systems (EFAS) currently issues warning flood notifications based on static metrics. This study proposes a data-driven alternative of notification issuance, to determine whether a supervised machine learning model can improve the accuracy of EFAS's decision-making process. A Multi-layer Perceptron Artificial Neural Network was developed and trained on historical EFAS forecasts and observed discharges from four Dutch river stations. The model's performance was compared against a surrogate of the current EFAS v5.2 operational logic. Results indicate that the machine learning model outperforms the current system for short- to medium-term forecasts (2-4 days horizon), highlighting the potential of data-driven methods.

Table of Contents

1. Introduction

1.1. Background

Among natural hazards, floods are recognised as one of the most destructive and costly disasters worldwide (Wallemacq & House, 2018; CRED, 2020) and are generally defined as the overflow of water onto land that is normally dry (Kundzewicz et al., 2013). Although they vary in their magnitude and impact, they are commonly perceived by the public as events that cause significant disruption, damage, and loss of life (Mishra et al., 2022). Floods can result from a range of causes, including intense rainfalls, storm surges, landslides, rapid snowmelt, high tides, or failure of man-made structures (Kundzewicz et al., 2013). These causes are often interconnected and can be intensified by compounding effects, meaning events where multiple hazards happen simultaneously or sequentially, both in space and time (Seneviratne et al., 2019). Examples include the combination of wind and rain with high tides (Thelen et al., 2024) or saturated soils followed by intense rainfall events (Seneviratne et al., 2019). Moreover, these interactions are complex and dynamic, as they are subject to change over time in response to climate change, as well as land-use changes such as deforestation and urbanisation (Mishra et al., 2022).

Within the various types of floods, fluvial floods occur when a river has a rapid rise in water levels inundating adjacent areas and are often followed by a slower decline in discharge (Doocy et al., 2013). These floods are alarming due to their possible impact on large areas along a river system. Moreover, growing evidence are proving that climate change is aggravating the frequency and the intensity of extreme river discharge events. One significant consequence of this trend is the shortening of the expected interval between major flood events, meaning that what was once considered rare '100-year' floods are now occurring more often (Alfieri et al., 2015; Blöschl et al., 2019).

The Intergovernmental Panel on Climate Change (IPCC) and the 2024 European State of the Climate report both identify Europe as one of the regions with the highest projected increase in surface water flood risk (Biesbroek et al., 2022; C3S & WMO, 2025). Over the past few years, Europe has experienced numerous major floods that have frequently made headlines. For example, the devastating floods in Germany and the Benelux region in July 2021 resulted in approximately 200 fatalities, damaged 72,000 buildings and €50 billion in damages (EEA, 2025). More recently, in September 2024, widespread flooding affected eight countries across Eastern Europe (Jones, 2024) and in October 2024, severe storms in Valencia, Spain, caused river overflows that led to catastrophic damage and approximately 150 deaths (Henley & Jones, 2024). According to the European State of the Climate report, floods in 2024 alone resulted in the deaths of at least 335 people and affected an estimated 413,000 individuals (C3S & WMO, 2025). Moreover, the PESETA IV study, which projected the economic impacts due to climate change by the European Commission's Joint Research Centre (JRC), assessed that currently

river flooding causes approximately €7.8 billion in damages annually in the EU and the UK (Feyen et al., 2020). Without mitigation and adaptation measures, these numbers could rise significantly by the end of the century.

Therefore, effective flood management has become an important priority in response to the increasing flood risks. To achieve meaningful disaster risk reduction, it needs a comprehensive approach that merges various interventions, typically categorised into structural and non-structural measures. Structural measures refer to physical and engineer-based interventions designed to reduce the impact of flooding by enhancing resistance and resilience in structures (UNDRR, 2017). These include the construction of physical defences that control and redirect water, such as dikes, levees, flood barriers and drainage systems. In contrast, non-structural measures focus on reducing vulnerability and increasing societal preparedness without altering the physical environment. These measures include land-use planning, legislation, public awareness campaigns and educational programs (UNDRR, 2017). Among the non-structural approaches, one of the key components is the Early Warning System (EWS), which, according to the United Nations Office for Disaster Risk Reduction (UNDRR, 2016), refers to an integrated system that involves monitoring, forecasting, assessing the risk, communicating and preparing for it. The purpose of an EWS is to enable individuals, communities, and governments to act timely to reduce disaster risks before a hazard occurs. While the concept of EWS can be broad, internationally recognised systems are expected to include four essential components: risk knowledge, monitoring and warning, dissemination and communication, and response capabilities (Šakić Trogrlić et al., 2022). When fully implemented, these elements work together to enhance preparedness.

Within this framework, the European Flood Awareness System (EFAS) was introduced and established in 2012 by the JRC to strengthen the flood disaster response across Europe. It was developed to produce, as one of their products, medium-range streamflow forecasts and early warning information on riverine floods up to 15 days in advance (Smith et al., 2016). EFAS operates by using an ensemble weather forecast to drive a process-based hydrological model (LISFLOOD) and then simulate river discharges at high temporal and spatial resolution (Smith et al., 2016). The system consists of a structured organisation of specialised centres, each responsible for distinct operational functions, such as collecting hydrological and meteorological data, computing forecasts, and disseminating them (Smith et al., 2016). The dissemination centre issues forecast twice daily, along with other flood-related products, to EFAS partners, which include European and national authorities. In the case of flood alerts, notifications are also sent to relevant civil protection services. Currently, EFAS issues formal flood notifications based on a metric: a notification is triggered when at least 50% of the ensemble forecasts predict that river discharge will exceed the local 5-year return level within a window period of 2 to 7 days (O'Regan, 2024b).

1.2. Problem Statement

While floods are inherently unpredictable regarding their scale and timing (Kumar et al., 2025), warning information is expected to be both timely and accurate. However, EWS for floods inevitably face inherent limitations coming from input noise in meteorological and hydrological data, imperfect resolution, and the intrinsic uncertainty of model physics (Li et al., 2021). As a result, post-processing techniques are just as important as pre-processing and model improvements when it comes to improving the reliability of a forecast. Translating probabilistic forecasts into actionable alerts requires carefully balancing sensitivity, meaning the ability to detect actual flood events, against specificity, the ability to avoid false alarms. Increasing sensitivity by issuing more frequent warnings may reduce missed alarms on floods, but it also increases false positives (LeClerc & Joslyn, 2015). This risk leads to a *'cry wolf'* effect, in which repeated false alarms decrease the trust among the vulnerable to the risk. This trade-off shows the importance of continuing to improve both forecast skills and decision rules to ensure that EWS remains accurate and societally effective.

Within this context, EFAS plays a central role as Europe's continental EWS for fluvial floods. Since it was created, EFAS's forecasting module has demonstrated steady improvements in performance, measured through correct alerts, false alarms and misses compared to observed discharge data (Smith et al., 2016). Recent bulletins report that approximately 65% of EFAS's formal notifications corresponded to observed flood events. However, this figure is based on received feedback for only 23% of the disseminated formal notifications (EFAS, 2024), suggesting uncertainty in the true performance level.

While recent advances in machine learning (ML) have led to significant improvements in weather and hydrological forecasting capabilities (e.g. ECMWF's Anemoi project and the high-resolution AI Forecast System) (Maskell, 2023; Lentze, 2025), less attention has been given to ML application in the decision-making components. In EFAS, the decision to issue a formal notification remains based on a fixed-threshold approach, which does not benefit from the opportunities of data-driven approaches. To date, no published studies have directly integrated ML into EFAS's decision rules for issuing flood warnings, highlighting a clear research gap. To address this knowledge gap, a supervised ML approach will be proposed as an alternative to the current static metric. Supervised means that a model is trained on historical data with known outcomes to be able to learn patterns and correctly predict the outputs. This approach is supported by recent studies that use supervised ML techniques to enhance similar decision-making tasks. For instance, Muñox et al. (2021) used supervised learning algorithms to classify the conditions of a river in Ecuador into three decision-making categories: No-alert, Pre-alert and Alert. Their model showed high accuracy in predicting the appropriate alert level.

2. Objective and Research Questions

The main objective of this research is to focus on the monitoring and warning component of EFAS by proposing and evaluating a ML approach to improve the decision-making process within its flood notification system. Specifically, the study aims to design a supervised ML-based model that can more accurately determine when flood alerts should be issued, thereby improving the accuracy of EFAS's decision rule to support flood preparedness and response. As a result, the main research question addressed in this thesis is:

How can a supervised machine learning model improve the accuracy of EFAS's decision-making process for issuing flood warning notifications?

To answer the main research question, the research is structured into three sub-questions (SQ) as follows:

1. *How does EFAS's decision-making process for issuing flood notifications currently operate?*
2. *How accurate has the EFAS's flood notifications system been in recent years, based on historical forecast and observation data?*
3. *How accurately can a supervised machine learning model predict flood notifications, and how does its performance compare to the current operational system?*

3. Theoretical and Conceptual Framework

3.1. EWS Components

In 2022, United Nations Secretary-General António Guterres declared that all people on Earth must be protected by an EWS within five years (WMO, 2022). Already, the Sendai Framework for Disaster Risk Reduction 2015-2030 has a key target to reduce disaster losses, which has led to the launch of the *'Early Warnings for All'* program designed to strengthen global early warning coverage. Therefore, it is important to state what an effective EWS means. According to Šakić Trogrlić et al. (2022), it consists of four key pillars, shown in **Figure 1**.

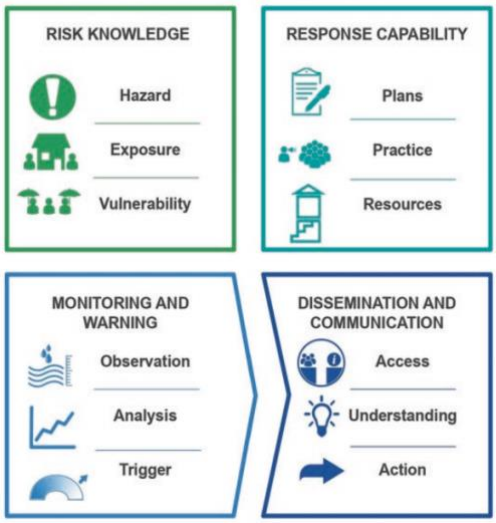

Figure 1. *Components of an EWS (Adapted from Šakić Trogrlić et al. 2022)*

Pillar 1: Disaster Risk Knowledge

The first element of EWS is understanding the disaster risk, which allows to make informed decisions (Šakić Trogrlić et al., 2022). This process depends on accessing relevant data about the hazard, evaluating possible impacts, and sharing the gathered information with all stakeholders (Šakić Trogrlić et al., 2022). In the context of fluvial flooding, risk identification must go beyond recognising flood-prone areas. It should also involve an understanding of the primary drivers, such as rainfall, weather patterns, soil moisture levels, river basin characteristics and snowmelt dynamics. Additionally, a risk assessment must be conducted to evaluate the threat's level and its social and environmental consequences.

Pillar 2: Hazard Detection, Monitoring, and Forecasting

The second pillar focuses on the monitoring and forecasting of hazards (Šakić Trogrlić et al., 2022). It is achieved by integrating real-time observations, historical data, and scientific modelling. Forecasting capabilities vary based on the type of hazard. For instance, tornadoes can only be predicted

minutes in advance, droughts have seasonal timescales, and the development of storms can be projected days ahead (Šakić Trogrlić et al., 2022). In the context of fluvial flooding, EFAS offers forecasts with a horizon of up to 15 days (Smith et al., 2016).

Pillar 3: Dissemination and Communication

Monitoring and forecasting are not enough, as the information must be communicated clearly and effectively to those at risk. Dissemination processes should ensure that warnings are delivered through accessible channels and presented in a format that is easily understood by diverse audiences (Šakić Trogrlić et al., 2022). While the strategy for communicating them often depends on the specific type of hazard, a major challenge of addressing the uncertainties in forecasts remains (Doyle et al., 2019). Nevertheless, education and trust in institutions are key to guaranteeing that warnings are not only received but also acted upon in a timely and appropriate manner.

Pillar 4: Response Capability

Finally, an effective response depends on the capacity of individuals, communities, and institutions to act promptly when receiving a warning (Šakić Trogrlić et al., 2022). Preparedness measures should be in place before a hazard occurs, and response strategies must also include actions taken during and after the event. Since floods and other hazards can cross administrative boundaries, transnational collaboration is critical. EFAS was created to facilitate such cross-border coordination, providing a pan-European platform for flood forecasting.

3.2. Improving EWS through Machine Learning

Current progress in research has significantly improved the ability of modern EWS (Šakić Trogrlić et al., 2022). Focusing on the second pillar, *"detecting, monitoring, and forecasting hazards"*, with higher access to open-source tools and growing accessibility of environmental data, data-driven alternatives to traditional modelling approaches have become increasingly available. While physically based models require extensive expertise and intensive computation, data-driven models treat the system as a *'black box'* and use statistical or ML techniques to map inputs to outputs (Mosavi et al., 2018). They depend exclusively on large volumes of historical data to learn input-output relationships without requiring knowledge of the underlying physical processes (Fares et al., 2023). Therefore, modern ML methods, a branch of Artificial Intelligence (AI), are well-suited to processing high-dimensional, nonlinear, large, and complex datasets, aligning effectively with the current 'big data' driven environmental monitoring. They are tools capable of identifying nonlinear patterns and delivering rapid and accurate predictions (Chamola et al., 2020; Khan et al., 2023). Hence, since 2011, there has been a shift toward ML methods becoming increasingly dominant in flood forecasting applications (Byaruhanga et al., 2024).

As there is currently a large variety of possible ML models based on the type of dataset and prediction task, they are broadly categorised into four main types: supervised, unsupervised, semi-

supervised, and reinforcement learning (Kumar et al., 2025). Supervised learning has a dominant role in flood forecasting, where models are trained on labelled data to predict either continuous (e.g., river discharge) or categorical (e.g., flood severity classes) outcomes. Common supervised learning algorithms include regression models that estimate relationships between input and output variables (Hidayat Jati et al., 2019), decision tree-based methods such as Random Forest and gradient-boosted trees (XGBoost) (Muñoz et al., 2018), and Support Vector Machines (SVMs) (Wu et al., 2019; Yu et al., 2017). Additionally, Neural Networks, including deep architectures, are increasingly used for their capacity to model complex patterns within very large-scale datasets (Muñoz et al., 2021; Kumar et al., 2025). In particular, recurrent networks such as Long Short-Term Memory (LSTM) are often used to model temporal dependencies in time-series data (Nevo et al., 2022), while Convolutional Neural Networks (CNNs) are used to capture spatial patterns in maps or remote sensing images (Bentivoglio et al., 2021).

Although supervised learning algorithms are more popular in EWS applications and generally more suitable for prediction tasks (Kumar et al., 2025), unsupervised ML algorithms are gaining popularity, particularly in studies where ground-truth data is scarce. As these models do not rely on labelled data to detect hidden patterns, they can be used in flood detection studies in environments with limited data (Tanim et al., 2022). Semi-supervised learning is meant for complex problems with a limited amount of labelled data and a large amount of unlabelled data to improve model performance while reducing the need for extensive manual labelling (Kumar et al., 2025). Finally, reinforcement learning focuses on developing an optimal policy that can map states to actions in a dynamic environment and has been explored for rescue path planning (Li et al., 2023).

3.3. EFAS's Operational Workflow and Research Positioning

To understand how ML could enhance EFAS, it is first important to know how EFAS currently operates. Therefore, a conceptual framework will first be developed to outline EFAS's operational workflow, followed by a delineation of the research within that workflow. Established under the Copernicus Emergency Management Service (CEMS), EFAS provides continental-scale probabilistic flood forecasts and early notifications to national and regional water management authorities (Adams & Pagano, 2016). The system relies on a wide range of input data to achieve this, which includes meteorological and hydrological observations from in-situ measurement stations and radar networks, used to define the initial conditions of the hydrological model. A network of national meteorological and hydrological services supplies real-time and historical data.

For the meteorological forecast input data, EFAS relies on different meteorological model products, each differing in resolution and forecast range (Smith et al., 2016). An overview of these models is

provided in **Table 1**. Two are based on the European Centre for Medium-Range Weather Forecasts (ECMWF) integrated forecasting system, which provides a control high-resolution run called ECMWF-HRES, capable of predicting up to 10 days, and ECMWF-ENS, an ensemble of 51 perturbed forecasts of lower resolutions that can provide medium-range probabilistic forecasts up to 15 days ahead (Smith et al., 2016). Ensemble refers to a set of forecasts from multiple simulations, in which each run has variations in its initial conditions. Hence, the ECMWF-ENS is essential to assess uncertainty by presenting a range of possible future weather outcomes throughout the forecast execution (ECMWF, 2020; Owens & Hewson, 2018). It allows us to assess the predictability of the atmosphere by examining how closely it aligns with the control member forecast (ECMWF-HRES). The Deutscher Wetterdienst (DWD) provides a further deterministic forecast with high spatial resolution, produced by combining ICON-EU and ICON models (O'Regan, 2025). However, its forecast range is limited to up to 7 days (Smith et al., 2016). To focus more on identifying high-impact local weather events, EFAS implements a 20-member ensemble from the COnsortium for Small-scale MOdelling Limited-area Ensamble Prediction System (COSMO-LEPS). The range is shorter, as it provides forecasts only 5 days ahead (Smith et al., 2016).

Table 1

Overview of EFAS meteorological forecast systems. The spatial resolution is from O'Regan, (2025).

Forecast name	Members	Maximum horizon	Spatial resolution
ECMWF-HRES	1	10 days	~ 9 km
ECMWF-ENS	51	15 days	~ 18 km until 26-06-2023 ~ 9 km from 27-06-2023
DWD	1	7 days	~ 13 km
COSMO-LEPS	20	5 days	~ 7 km

All four models generate forecasts daily and twice a day (at 00:00 and 12:00 UTC), providing an extensive dataset that captures evolving weather patterns (Smith et al., 2016). These results are then forced into the hydrological rainfall-runoff routing model LISFLOOD, which is further integrated with real-time data, satellite observations and remote sensing data (De Roo et al., 2000). LISFLOOD, developed by the JRC, is a physically based model designed for operational flood forecasting at a European scale. It simulates a complete water balance at a 6-hour time step and for every 1,5 km grid cell, producing an ensemble of hydrological forecasts (LISFLOOD version 5). EFAS hydrological simulations, like the meteorological forecasts, are updated twice daily.

To calibrate the forecasts, EFAS applies a post-processing procedure based on two main approaches: offline and online calibration (Matthews et al., 2022). Offline calibration is conducted twice a year, to establish the statistical distributions of observed and simulated river discharges, using techniques like

kernel density estimation and the generalised Pareto distribution for extremes (Matthews et al., 2022). These distributions are then mapped into a joint distribution estimation to be able to understand the expected future observations. Online calibration is performed daily at each station, updating forecasts using the most recent observations. First, the offline calibration adjusts the most recent discharge observations to correct for systematic errors, then the ensemble forecast is transformed into a probability distribution. Finally, both outputs are merged using an Ensemble Kalman filter to produce the final probabilistic forecast (Matthews et al., 2022). The probabilistic forecast means that rather than providing a single predicted value, it estimates the probability of possible future river discharges based on the information available up to date (Todini, 2008).

This post-processing procedure results in the EFAS's first product for its partners: station-level post-processed probabilistic hydrographs. Based on these forecasts, EFAS offers a range of products to its partners, including formal flood notifications via email and on the secure web portal, providing access to the most up-to-date information (Smith et al., 2016). To standardise decision-making across different stations, as EFAS does not have the local knowledge at each forecast point, it applies a simplified binary flood classification (Pozo et al., 2015). Specifically, EFAS evaluates whether forecasted streamflow exceeds predefined flood thresholds derived from return periods, which represent the average interval between flood events of a given magnitude. The return periods used in EFAS are 1.5 years (low), 2 years (medium), 5 years (high), and 20 years (severe) (Pozo et al., 2015). These thresholds are calculated for each river based on historical simulated discharge data, called simulation forced with observed (SFO) (EFAS, 2024), and are used as the benchmark for flood classification. In this research, I will adopt the same definition of flood.

Overall, **Figure 2** provides an overview of EFAS's operational workflow. This research is positioned within the framework, aiming to improve the segment between the post-processing stage and the dissemination of formal notifications to EFAS partners.

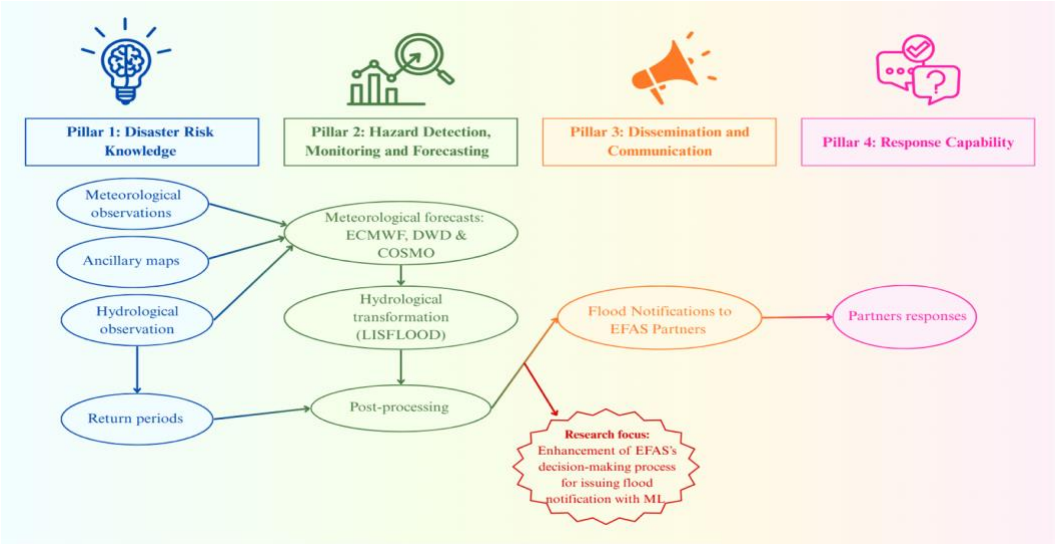

Figure 2. *The conceptual and theoretical framework of EFAS is placed within the four pillars of EWS. The research focus is highlighted within the diagram to show its specific position in the overall EFAS workflow*.

4. Study Area and Data

4.1. Study Area

The study focuses on four river gauge stations in the Netherlands: Lobith, Megen Dorp, Venlo and St Peter Noord in Maastricht. These locations were selected as they are EFAS hydrological model performance points, where EFAS LISFLOOD hydrological simulations are compared against observed discharge records, as part of the offline calibration process. (CEMS, 2019; O'Regan, 2023). These four stations have available long and continuous observed discharge datasets, offering a reliable dataset for evaluating both regular and high-flow events. Lobith (51.00° N, 6.11° E), a city in the Netherlands located on the German-Dutch border, where the Rhine enters the Netherlands and divides into two main tributaries, the Pannerden Canal and the Waal, as seen in **Figure 3**. Due to its strategic location, Lobith plays a central role in national flood in flood risk management and water resource planning (Rijkswaterstaat, 2024). Therefore, Lobith has one of the most extensive databases in Europe, with daily discharge measurements since 1866 (Bomers et al., 2019). Megen Dorp (51.82° N, 5.4° E), Venlo (51.56° N, 6.05° E), and St. Pieter Noord (50.8° N, 5.7° E), are three-gauge stations located along the Meuse, a major rain-fed river entering the Netherlands from Belgium. Both rivers' water levels throughout the year are closely monitored by the Dutch national water authority and Rijkswaterstaat.

This study uses two primary data sets: the hydrological forecast of river flow from EFAS (Section 4.2) and observed discharge data from Dutch authorities (Section 4.3). Together, these datasets will help answer the main research question.

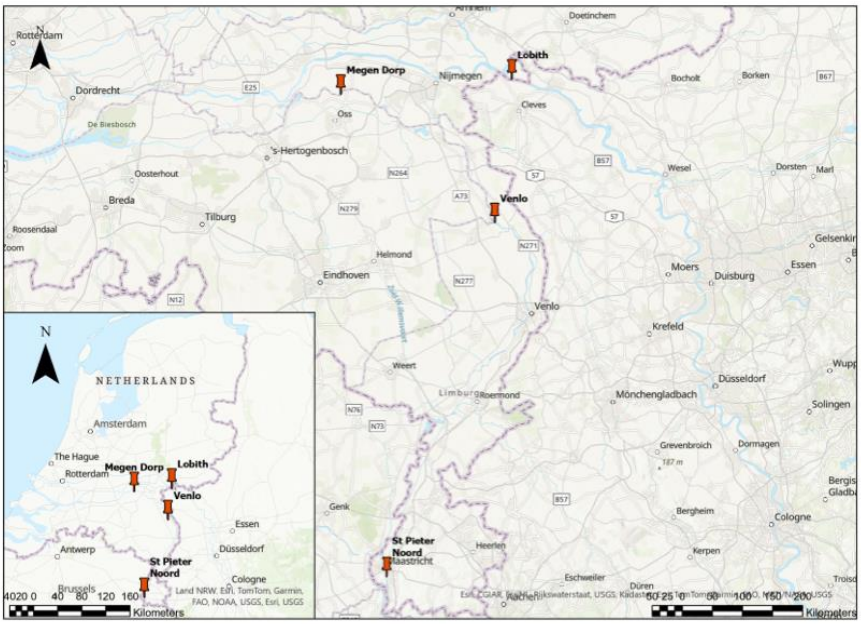

Figure 3. *Map showing the locations of the four gauge stations (Lobith, Venlo, Megen Dorp and St Peter in Maastricht) within their regional and national context.*

4.2. EFAS Forecast River Discharge Data

The EFAS forecast river discharge data is a core product and represents the volume rate of water flow, including water, sediments, chemical, and biological materials, through a river channel cross-section. Discharge values represent the hydrological forecasts generated using the LISFLOOD hydrological model driven by meteorological predictions from ECMWF, COSMO and DWD (see Section 3.3). The dataset is sourced from the JRC & CEMS (2019) and covers the period from October 10th 2018 to present. Real-time access is restricted only to EFAS partners, while for external users, the data are released with a 30-day delay relative to the current date.

Each EFAS forecast provides discharge estimates over a rolling prediction window of 360 hours (15 day) with a 6-hour resolution. This results in 60 forecasted discharge values per initialisation of the forecast. As stated in Section 3.3, the ECMWF-ENS consists of 51 members: one control member (ECMWF-CON) with unperturbed initial conditions and 50 perturbed ensemble members. Due to the configuration of the CEMS server, the control member is handed separately when downloading. As new meteorological data becomes available, the rolling window is updated to generate the next forecast cycle with the latest meteorological input. Given that official EFAS flood notifications are issued for the period between day 2 and day 7 forecast, only discharge values corresponding to horizons between hour 48 and hour 168 are used in this study. **Table 2** shows an overview of the available forecast datasets, including their temporal resolutions and periods of missing data. The data are unavailable due to an error in the calculation of the initial conditions for EFAS version 5 (O'Regan, 2024c).

Table 2

Overview for each dataset, its resolution window, the available period and the period with missing data.

Dataset	Resolution Window	Available Period	Missing Data
ECMWF-ENS	6-hourly	10/2020 – 04/2025	19/09/2023 - 29/05/2024
ECMWF-CON	6-hourly	10/2020 – 04/2025	19/09/2023 - 29/05/2024
ECMWF-HRES	6-hourly	10/2018 – 04/2025	19/09/2023 - 29/05/2024
DWD-HRES	6-hourly	10/2018 – 04/2025	19/09/2023 - 29/05/2024
COSMO-LEPS	6-hourly	01/2019 – 04/2025	19/09/2023 - 29/05/2024

Discharge data on Version 5.0 of the EFAS model are provided as a gridded time series on a regular latitude-longitude projection. The dataset offers pan-European coverage, but for this research, an area centred on all of the river gauges will be extracted as a sub-region. The spatial resolution of the forecast data is approximately 1×1 arcminute, equivalent to roughly 1.5 × 1.5 kilometres (JRC & CEMS, 2019). While river discharge is available at the surface level, the model also includes other variables such as

soil moisture across three vertical layers and snow depth. Finally, the dataset can be formatted in both GRIB2 and NetCDF-4 formats. For efficient and automated access, data retrieval is conducted via Python-based application programming interface (API) provided by CEMS. Deterministic data from DWD-HRES and ECMWF-HRES, and the control member ECMWF-CON, are downloaded via a Python-based application programming interface (API) provided by CEMS on a HPC system, following their best-practice guidelines (O'Regan, 2024a). Meanwhile, the ensemble data are obtained through the web server, as it was faster. After the download, the raw files are processed and converted into CSV for further analysis. Details of the data preprocessing are further explained in Appendix A.

4.3. Observed River Discharge Data

Observed river discharge data used in this study are sourced from '*Waterinfo Extra'*, a platform managed by Rijkswaterstaat, the Dutch national water authority (Rijkswaterstaat, 2025). This website provides a variety of in-situ measurements from various monitoring stations across the Netherlands. For the purpose of this research, the relevant dataset to be used is the river flow rate in m^3/s, which captures the actual volume of water passing through a cross-section of the river, recorded at regular intervals. The discharge data is measured every 15 minutes, however, for consistency with the EFAS forecast dataset, it will be aggregated to a 6-hour resolution using the mean. Moreover, to maintain consistency and to have a meaningful comparison, the temporal coverage of the observed data will be aligned with that of EFAS. This is important, as the observed discharge is the ground truth in this research, to check whether the river's observed discharge exceeds the return period threshold. In addition, historical discharge data from 1990 were downloaded for each station to compute the return period.

On the website, users can request the data download, and a link to a CSV file will be sent via email. However, during data collection, '*Waterinfo Extra'* underwent a system-wide synchronization process that temporarily limited the access to the 2024 measurements via the Web Services. Therefore, 2024 data was retrieved via the Waterinfo Extra API. Due to this synchronization, the data varies slightly between stations. Lobith and St. Peter Noord have data until April 2025, while Megen Dorp and Venlo until the 31st of December 2024.

5. Methods

This chapter will outline for each sub-research question the methods that were used to answer it. **Figure 4** visually presents a general overview of the methodological workflow.

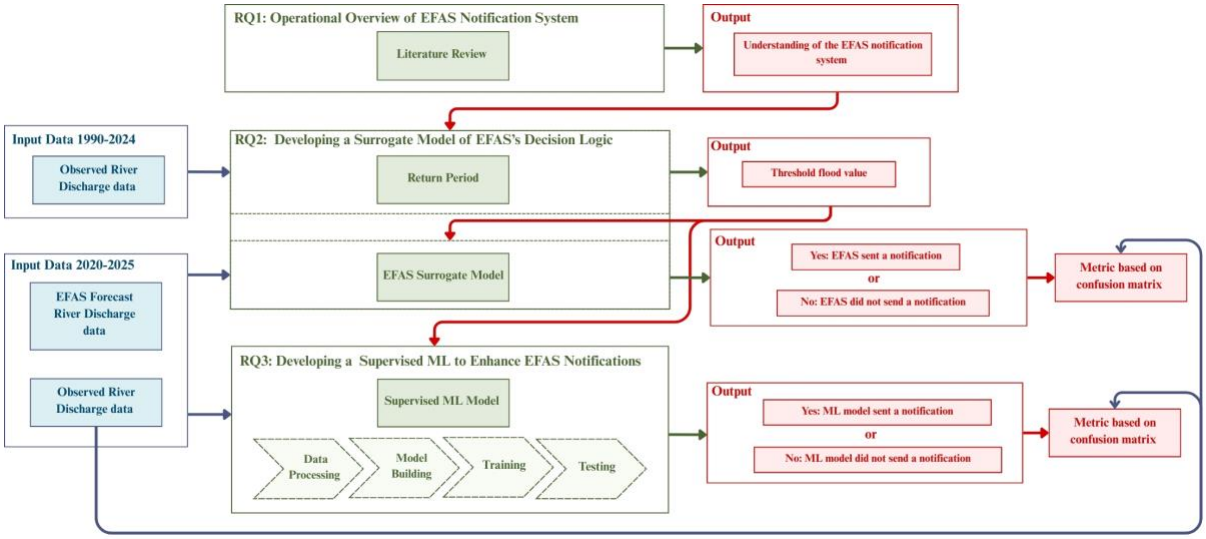

Figure 4. *Overview of the methodological workflow. The blue box show the input data used, the green boxes represent the research questions together with the methods used to address them, and the red boxes display the outputs from each model. The red arrows show how outputs from one module serve as inputs to the next module.*

5.1. Operational Overview of EFAS Notification System

This section outlines the methods used to answer the first research question, *"How does EFAS's decision-making process for issuing flood notifications currently operate?"*. The objective was to reconstruct EFAS's notification system through a literature review. The analysis focused on academic and peer-reviewed literature discussing or evaluating EFAS, as well as on technical reports and grey literature (Wiki pages, user guides, and bulletins) published by EFAS, ECMWF, or JRC, which often provide descriptions absent in journal articles. Relevant documents were identified using Google Scholar, Scopus and official institutional repositories, filtering for titles, keywords or content with terms such as *"EFAS Notification*" or *"EFAS Alert"*. Data management and thematic analysis were done on ATLAS.ti (ATLAS.ti Scientific Software Development GmbH, 2023). Each document was coded thematically using an inductive-deductive approach, meaning initial codes were established based on known EFAS components that I was looking for, while additional codes were added during the analysis to include new relevant information. The eight thematic code groups are:

1. Forecast Inputs: identification of model providers and ensemble structures.
2. Aggregation Logic: methodology for combining different forecasts.
3. Formal Notification: general information about the rules of formal notifications.
4. Operational Constraints: spatial and temporal conditions limiting the system.

5. Evaluation Methods: performance metrics used to assess model configurations.
6. System Evolution: chronological development of EFAS versions and criteria updates.

The outcome of this analysis will describe a synthesis of the operational overview of the EFAS decision-making chain, derived from the reviewed materials.

5.2. Developing a Surrogate Model of EFAS's Decision Logic

From the overview of the first research question, the second research question, *"How accurate has the EFAS's flood notifications system been in recent years, based on historical forecast and observation data?",* aims at developing a surrogate model to replicate the operational logic of EFAS's current flood notification process. First the threshold used to classify events as flood or non-flood conditions are calculated per station. Then, a model is designed to approximate the decision rules used by EFAS to assess the historical accuracy and provide both a quantitative baseline for evaluating EFAS's current system and a feature for training the ML.

5.2.1 Threshold of Flood Return Period

The EFAS notification system converts forecasted discharge time-series into binary classification of exceedance or non-exceedance based on a predefined discharge threshold (Casado-Rodríguez et al., 2025). These thresholds correspond to the 5-year return period (Q_5) of river discharge, derived by fitting a Gumbel distribution to the annual maximum values of the river from 1990 to the current year, a common method in hydrology to model extreme events. (Casado-Rodríguez et al., 2025). The discharge value (x_T) associated with a return period T (years) is obtained using Equation (1).

$$X_T = \mu - \beta \ln\left[-\ln\left(1 - \frac{1}{T}\right)\right] \quad (1)$$

where μ represents the average flood magnitude and β represents the variability of the flood, scaled by a frequency factor reflecting the event's rarity (Anghel, 2024; Jonsson & Rydén, 2017). The parameters of the Gumbel distribution were estimated using the maximum-likelihood method from the SciPy library.

EFAS, however, does not use real observed discharge data for this computation. Instead, it relies on long-term discharge simulation produced by LISFLOOD hydrological model forced with gridded observed discharge, precipitation and temperature data (Mazzetti et al., 2023). However, in this study the return period is derived from observed river discharge data at each gauge station, obtained from Waterinfo Extra, ensuring that is grounded on actual hydrological behaviour. It allows locally representative thresholds while reducing bias related to EFAS model simulations.

5.2.2. Accuracy Metrics

To evaluate the surrogate model, predictions are compared against observed discharge events. Standard accuracy measures the proportion of correct classifications by computing the number of correct predictions divided by the total number of predictions, to give the right proportion of correct predictions. However, this metric alone is insufficient for evaluating a classification model (Fergus & Chalmers, 2022), as it treats all errors equally. Therefore, to provide a more class-sensitive evaluation, additional metrics were used to assess the accuracy of the model. These rely on the confusion matrix (**Table 3**), a 2x2 table that compares predicted values with actual outcomes into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (Hossin & Sulaiman, 2015; Naidu et al., 2023). Each type of misclassification has different operational implications. FN, meaning no alert was issued but a flood happened, represents the most critical error from the EWS perspective, as it represents a failure to provide timely protection. FPs, where an alert was given but no flood occurred, are less severe but they may lead to alarm fatigue and affect the trust in the system (the *"cry-wolf"* effect).

Table 3

Overview of the Confusion Matrix Table

	Observed Flood (1)	**Observed No Flood (0)**
Predicted Flood (1)	True positive (TP)	False Positive (FP)
Predicted No Flood (0)	False Negative (FN)	True Negative (TN)

To provide class-sensitive assessments, metrics such as precision, recall and the F1 and Fβ Score are used (Chicco & Jurman, 2023), which provide the ratio of the model's ability to correctly identify positives and negatives. **Table 4** provides all the formulas and descriptions of each selected metric. The F1-score, a harmonic mean of precision and recall, summarizes the trade-off between false positives and false negatives at a fixed threshold (Hossin & Sulaiman, 2015; Naidu et al., 2023). Unlike the F1-score, which weights precision and recall equally, the Fβ allows for prioritizing one metric over the other. The specific value of β is selected based on the findings from the literature review regarding EFAS's operational decision-making process (SQ1).

Finally, the Area Under the Precision-Recall Curve (PRAUC) evaluates the model's performance by showing the trade-off between precision and recall across all possible thresholds (Sofaer et al., 2019). Unlike the other metrics derived directly from a confusion matrix, PRAUC is computed by plotting precision and recall ratio at different threshold values and calculating the area under that curve, so it does not have a single fixed formula. As it relies only on TP, FP, and FN, and ignores TN, PRAUC is less affected by class imbalance and by the infrequent occurrence of the positive class (Sofaer et al., 2019).

Table 4

Overview of the selected metrics to assess the model performance.

Metric name	Formula	Description
Precision	$\frac{TP}{TP + FP}$	Portion of predicted flood events that were actually floods.
Recall	$\frac{TP}{TP + FN}$	Ability to correctly identify actual flood events.
F1-score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	Harmonic mean between recall and precision values.
Fβ-score	$(1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$	A weighted harmonic mean where $\beta < 1$ prioritizes precision, while $\beta > 1$ prioritizes recall.
PRAUC		Trade-off of how well the model separates flood from non-flood events across all possible classification thresholds.

5.3 Design of Supervised ML Model and Performance

This section describes the design of the supervised ML developed to address the third research question "*How accurately can a supervised machine learning model predict flood notifications, and how does its performance compare to the current operational system*?". To answer this question, this research uses neural networks as a supervised machine learning algorithm, to determine whether a flood notification should be issued.

Flood forecasting is a complex modelling challenge due to the non-stationary and non-linear nature of hydrological data in both time and space (Cloke & Pappenberger, 2008). Hydrological behaviours are rarely consistent over time, and the relationships between variables shift dynamically. Additionally, extreme events are infrequent, resulting in historical datasets with limited examples of critical conditions, which further increases the modelling difficulty (Cloke & Pappenberger, 2008). To adequately capture these dynamics, the model's complexity needs to align with the complexity of the forecasting task. Because of the nature of the problem, Neural Networks (NNs) are particularly well-suited for capturing complex non-linear relationships between inputs and outputs (Islam et al., 2019; Kumar et al., 2025).

5.3.1. Input Data Organization

The dataset is organized in a structured tabular format, with each row representing a single forecast instance defined by the date and forecast location. All predictor variables available at that instance are included within the same row. As detailed in Section 4.2, forecasts are issued at 00:00 and 12:00 UTC, resulting in two rows per day for each location. While the operational system outputs data every 6 hours, giving four forecasts per day, this analysis focuses on a single forecast horizon per day to reduce computational complexity. Specifically forecast horizons of 48, 72, 96, 120, 144, and 168 hours are analysed, corresponding to the final horizon within each day. For clarity, the horizons are hereafter referred to by day number (e.g. 48 hours as day 2). The input features primarily consist of discharge forecasts from the different provider systems, which are incorporated based on their configurations and structures. Deterministic forecasting systems (DWD, ECMWF-HRES, and ECMWF-CON) provide a single discharge value for each forecast horizon at six-hour resolution, which are included as individual numerical features. Ensemble-based systems (COSMO-LEPS and ECMWF-ENS) provide multiple forecasts per horizon. To summarise the ensemble information, the ensemble median is used for each system. This provides a robust measure of central tendency that is less affected by outliers in high-uncertainty scenarios. Additionally, the observed discharge measured twelve hours prior to forecast issuance is included to provide antecedent hydrological information. All numerical features were standardized using z-score scaling (mean of 0, standard deviation of 1) to ensure uniform contribution (Shyalika et al., 2024). The categorical feature “location” was transformed using one-hot encoding, converting it into binary vectors to prevent the model from assuming ordinal relationships between the river stations (Scikit-learn, 2019). Finally, the “Ground Truth” target variable is derived from the observed discharge dataset and, as the problem is formulated as a binary classification task, takes a value of 1 if the observed discharge exceeds the flood threshold, and 0 otherwise. While this variable is not used as an input feature, it plays an important role as the objective target for model training and as the reference standard for comparing the model’s probabilistic outputs, enabling the loss function to measure error and adjust network weights during training (Shrestha, 2024). **Table 5** provides an overview of the input features, including their descriptions, data types, and formats. During training, each dataset row is passed to the network as a vector of feature values, activating the corresponding input neurons.

Table 5.

Overview of the tabular dataset containing input features for the supervised ML model.

Feature Name	Description	Type	Format
Location	Identifier of the forecast river station: Lobith, Venlo, St. Peter, Megen	Categorical	One-Hot Encoder
DWD [h]	Deterministic discharge forecast from DWD at horizon *h* (6-hour resolution)	Numerical	One column per horizon
HRES [h]	Deterministic discharge forecast from ECMWF-HRES at horizon *h* (6-hour resolution)	Numerical	One column per horizon
CON [h]	Deterministic discharge forecast from ECMWF-CON at horizon *h* (6-hour resolution)	Numerical	One column per horizon
ENS median [h]	Median of all ECMWF-ENS ensemble members at horizon *h* (6-hour resolution)	Numerical	One column per horizon
COSMO median [h]	Median of all COSMO ensemble members at horizon *h* (6-hour resolution)	Numerical	One column per horizon
12-hour prior	Observed river discharge measured 12 hours prior to the forecast reference time	Numerical	Single column
Ground Truth [h]	Binary indicator of threshold exceedance (1 = flood, 0 = no flood)	Binary	One column per horizon

5.3.2. Model Architecture

Having defined the structure of the input features, the neural network architecture is introduced. A multi-layer perceptron artificial neural network (MLP-ANN) is used in this study, which is an architecture widely used for classification tasks across multiple complex domains (Tian et al., 2021). An MLP is a fully connected network, also known as a dense network, in which every neuron in one layer is connected to every neuron in the next layer (Islam et al., 2019). At its core, each neuron acts as a fundamental processing unit designed to detect specific patterns. It receives signals from the previous layer, assigns them varying levels of importance and aggregates them to determine its output. As **Figure 5** shows, the chosen model is strictly feedforward, meaning that information flows unidirectionally from the input layer, through the hidden layers, to the output layer. This design choice reflects the framing

of flood notifications as independent classification events based on the current input state, rather than as a sequential problem that would require a loop structure.

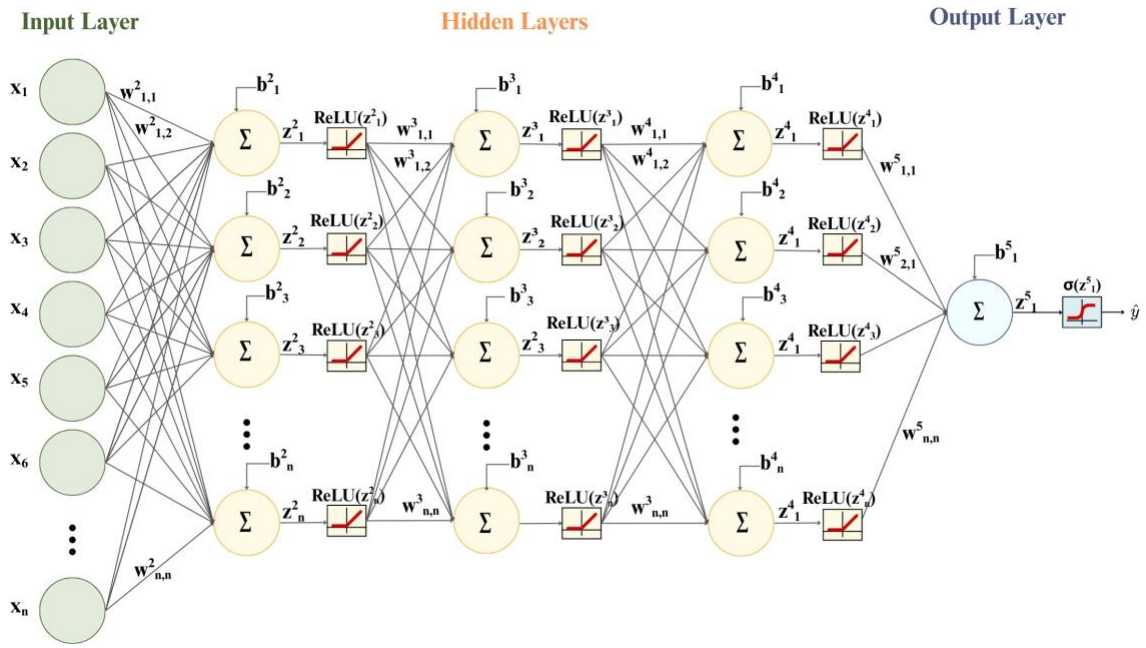

Figure 5. *Visualisation of the model architecture used in this study. `Each circle represents a neuron, and the network consists of an input layer, three fully connected hidden layers and single output neuron to generate the final flood notification probability ŷ.*

The model architecture is defined by several parameters that influence its performance. The input layer size corresponds to the number of input features in the dataset. Following the input layer, the network includes three additional layers of interconnected neurons, called the hidden layers. Each hidden layer is made of a defined number of neurons, which will be treated as hyperparameters and will be determined via tuning. Mathematically each neuron j computes the weighted sum of its inputs (x_i) plus a bias term (b_j) to generate a pre-activation value (z_j) (Equation 2).

$$z_j = \sum_{i=1}^{n} w_{ji} x_i + b_j \tag{2}$$

In this context, the weight parameter w_{ji} determines the strength of the connection between the input i on neuron j, while b_j allows the neuron to adjust its activation threshold. This linear combination is then passed through a Rectified Linear Unit ($ReLU$) activation function, which introduces non-linearity (Equation 3). Selected for its computational simplicity and effectiveness (Shrestha, 2024), ReLU outputs zero for negative inputs and returns the input value if positive.

$$ReLU(z_j) = \max(0, z_j) \tag{3}$$

The final layer has a single linear neuron that aggregates the features from the hidden layers. The model outputs a real-valued score z, which represents the probability for flood notification. To interpret

this score as a probability, the Sigmoid activation function (σ) is applied to the final pre-activation value z (Equation 4).

$$\sigma(z) \equiv \frac{1}{1 + e^{-z}} \tag{4}$$

This function transforms any real-valued input into an output $\hat{y}$ between 0 and 1. As the pre-activation value z increases, the output moves closer to 1, indicating a high confidence in a flood event. When z becomes very negative, the exponential term grows and moves the output toward 0 (Bishop, 1995; Nielsen, 2015). With Sigmoid function, small adjustments to the weights result in predictable changes in the output probability, preventing sudden prediction jumps. In this study, rather than explicitly applying the Sigmoid activation function, it is computationally fused with the Weighted Binary Cross-Entropy (WBCE) loss function (Equation 5) for higher stability (PyTorch, 2024). This function quantifies the difference between predicted outputs ($\hat{y}$) and the actual ground truth labels (y) to adjust the internal parameters to improve prediction accuracy over time (Shrestha, 2024). This formulation takes the value z as input and internally computes the log-probabilities.

$$L_{WBCE}(z, y) = -\left[\alpha \cdot y \cdot \log\left(\sigma(z)\right) + (1 - y) \cdot \log(1 - \sigma(z))\right] \tag{5}$$

Here y represents the ground truth, and α is positive class weight tuned for each forecast horizon to balance detection sensitivity and false alarm control (PyTorch, 2024). A high loss means poor predictive performance, whereas when low means alignment between the model's output and the observed outcomes. As the aim is to reduce this loss, the optimisation technique used is ADAM, an algorithm that provides extensions from the classic stochastic gradient algorithm and helps accelerate the performance of the model (Kingma & Ba, 2014).

MLP-ANNs have numerous tuneable parameters, which introduce inherent challenges related to the bias-variance trade-off. A model that is too simple fails to capture the underlying non-linear relationships in the hydrological data and has high bias, causing the model to underfit (Arbel et al., 2023). If the model is too complex, it captures noise rather than signals, leading to poor performance on unseen data and overfitting (Arbel et al., 2023). NNs tend to have low bias but are highly susceptible to high variance (Arbel et al., 2023). In this study, the scarcity of extreme events further intensifies this effect. When only a small part of the entire dataset is positive, a high-variance model may memorize these rare instances, rather than learning generalizable patterns. Therefore, targeted mitigation strategies are necessary to limit the model's variance while preserving the model's ability to detect complex patterns.

5.3.3. Training Configuration

The dataset consists of time-series observations, meaning a sequence of values in which time is an independent variable (Machiwal & Jha, 2012). Therefore, the data are temporally ordered and the model must preserve the sequential nature to function like an operational forecasting environment and prevent

temporal information leakage from future periods (Newaz et al., 2022). Due to the data gap from September 2023 to May 2024, data recorded before this gap is used for training and internal validation, while data after the gap serves as a hold-out test set (approximately 2087 rows, 20% of the total data). The training set consists of a set of data for which the correct outputs are known and can be used to train the model (Islam et al., 2019). To monitor the model's generalization performance and detect overfitting during the training phase, an internal validation strategy is required. To keep the temporal dependencies, a Time-Series Cross-Validation with an Expanding Window is employed (Hyndman & Athanasopoulos, 2018; Scikit-learn, n.d.). In this approach, the training set expands progressively over time, while the validation set always immediately follows the training period chronologically. As **Figure 6** shows, in the first fold the model trains on an initial segment of the data and validates on the following segment. In the next fold, the training window expands to include all data observed up to the end of the previous validation period, and validation is performed on the next unseen sequence. To ensure that each validation window contains enough flood events for meaningful evaluation, the number of folds was set to 2. This results in validation windows of circa 2780 rows each (~26.7% of the full dataset).

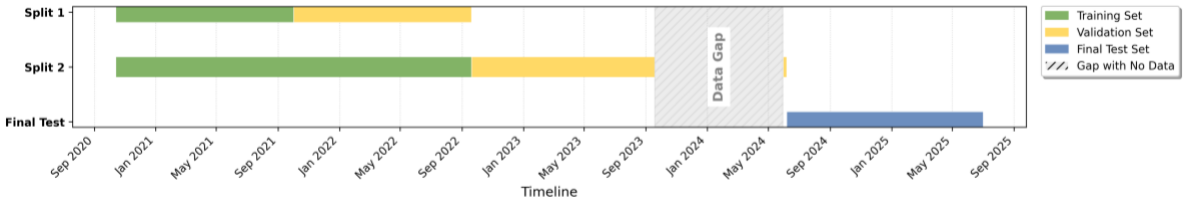

Figure 6. *Timeline of the data cross-validation structure into two expanding folds and a hold-out test period.*

5.3.4. Data Augmentation

When data is limited, model performance depends strongly on inductive bias, from model design and training procedures, to avoid overfitting, which data augmentation can increase (Arbel et al., 2023). After the splitting of the data, to deal with unbalanced datasets, data resampling is often used. However, standard resampling methods are often inappropriate for time-series data, as they destroy the dependence sequence (Härdle et al., 2003). To address this limitation, a targeted event-based oversampling strategy was applied to the minority class. Flood events are represented as contiguous temporal blocks B_k, defined as sequences of forecasts, during which the corresponding target variables are positive (Equation 6). By resampling the entire flood block rather than individual time steps, the augmentation process preserves the physical structure and persistence of flood, ensuring that the model is exposed to realistic flood dynamics.

$$B_k = \{t_m, t_{m+1}, \dots, t_p\} \text{ such that } y(t) = 1 \; \forall \, t \in [t_m, t_p] \tag{6}$$

Here t_m is the first observation in the tabular dataset for which the ground truth label equals 1, and t_p is the last contiguous observation with the same label. During training, these blocks are sampled with replacement and appended to the training dataset until a predefined ratio is achieved. This approach

increases the representation of flood patterns. In addition, to reduce the risk of the model overfitting, the study applies Gaussian Noise on the majority class during training (Panarin, 2024; Ye et al., 2023). Noise is added only to the forecast predictors, so the non-numerical fields and observations-based features are excluded. For each numerical non-flood training sample *i* and forecast feature *j*, a perturbed value is generated (Equation 7).

$$x'_{ij} = x_{ij} + \epsilon_{ij}, \qquad where\ \epsilon_{ij} \sim N(0, \sigma_j^2) \tag{7}$$

Here, $x_{ij}$is the original value and x'_{ij} the perturbed values and ϵ_{ij} is a random Gaussian random noise with 0 mean and feature-specific variance σ_j^2. Specifically, σ_j is tuned to ensure that perturbations remain within physically plausible ranges, preserving the interpretability of the data. These strategies are applied exclusively to the training set after temporal splitting within each fold. In practice, the effective size of the larger augmented training data is determined by the target class ratio between flood and non-flood samples, a tuneable hyperparameter. The validation and test sets remain unmodified to ensure a fair evaluation of real-world performance. The augmented dataset is then randomly shuffled before training, as the feedforward NN processes observations independently, and incorporated into an ensemble framework to further reduce variance (Fakhruzi, 2018; Ha et al., 2005). For each fold of the time-series cross-validation, three independent MLP models are trained on different stochastic realizations of the oversampled flood blocks and noise-augmented data. The probability outputs from these individual models are averaged to produce the final ensemble prediction.

5.3.5. Hyperparameter Tuning

The predictive performance and generalization ability of the model depend on a set of configuration parameters that define both the network architecture and the optimization process (Ilemobayo et al., 2024; Siadati, 2021). The hyperparameters were set up before the training, while the tuning process followed a manual iterative empirical approach. Initially, standard baseline configurations common in NN applications were applied, and then progressive adjustments were made based on the model's behaviour in validation, by monitoring bias-variance trade-offs through learning curves (Ilemobayo et al., 2024). Models with unstable loss or high fluctuations were interpreted as overfitting, while those with persistently high training loss were interpreted as underfitting. Confusion matrices were inspected for hyperparameters that directly influence the decision behaviour of the model, in order to assess FN and FP trade-offs. The hyperparameters tuned in this study can be categorized into different groups.

1. The specific input configuration was tuned depending on the prediction horizon to prevent overfitting while ensuring the ability to capture the most important information and relevant dynamics.
2. Model capacity includes the number of hidden layers and the number of neurons per layer that determine the network's ability to capture complex non-linear patterns (Ilemobayo et al., 2024).

3. Optimization is the process that adjusts internal weights and biases to minimize the loss function and uses the learning rate to determine the step size of the gradient updates, the batch size (the number of training sample used to estimate each gradient update) to balance computational efficiency with gradient stability, and the positive class weights (Ilemobayo et al., 2024).
4. Data augmentation is adjusted by the ratio of block sampling and the magnitude of the Gaussian noise (σ).
5. The choice of the threshold is also included. After all the ensemble models are trained and their prediction averaged, the model outputs a probability $\hat{p}(t) \in [0,1]$, representing the likelihood of a flood event at time *t*. To convert this continuous probability into a binary decision, a classification threshold (τ) must be chosen such that:

$$\hat{y}(\mathrm{t}) = \begin{cases} 1, & if\ \hat{p}(\mathrm{t}) > \tau, \\ 0, & otherwise \end{cases} \tag{8}$$

   The choice of threshold plays an important role in balancing FN and FP (Sheng & Ling, 2006). As such, the threshold is treated as a tuneable parameter. Several thresholds were evaluated during model development for each horizon to select a value that reflects an appropriate compromise between detection sensitivity and false alarm control.
6. Regularization is a set of techniques that constrain the model to prevent the network from fitting noise in the training data and reduce overfitting. In this study, Elastic Net regularization was applied, combining two parameters (Deepa et al., 2025; Farhadi et al., 2022; Siadati, 2021). First, L1 regularisation (λ_1) encourages sparsity by penalising large parameter values (w_i). Meanwhile, the other parameter L2 regularisation (λ_2) discourages large weights (w_i^2) more smoothly by shrinking them to 0 (Equation 9).

$$Elastic\ Net\ = L_{WBCE}\ +\ \lambda_1 \sum_i |w_i|\ + \lambda_2 \sum_i w_i^2 \tag{9}$$

   The hyperparameters λ_1 and λ_2 controlling the strength of these penalties, were optimized to minimize validation loss.

Overall, the final configuration of all tuned hyperparameters used in the trained model is reported in Table B1 in Appendix B. The same NN architecture and learning procedure are applied consistently across all forecast horizons. Differences between horizons arise only in hyperparameter tuning, reflecting changes in forecast uncertainty and information content.

5.3.6. Evaluation

The final hold-out test set is used only once, to provide a final performance assessment of the selected model, whose performance is benchmarked against the surrogate EFAS notification system, using the same evaluation metrics defined in SQ2 (see Section 5.2) to ensure consistency. This comparison assesses whether improvements are achieved over the current operational baseline. To

enable comparison across forecast horizons, each model's stochastic variability is controlled by fixing the random seed to a constant value of 1, which standardizes weight initialization and the shuffling order of training data. By maintaining these stochastic factors constant, any observed performance differences can be attributed to the information of the data rather than random initialization. All modelling tasks were conducted in Python using the PyTorch neural network library (Paszke et al., 2019) and scikit-learn (Pedregosa et al., 2011) for data preprocessing and metric calculation.

6. Results

6.1. Operational Overview of EFAS Notification System

To address SQ1 regarding the current operation of EFAS's decision-making process, this section provides details of the technical framework used to issue formal alerts. Although peer-reviewed literature on the specific operational dynamics of the EFAS notification system is limited, detailed technical documentation is available through official EFAS and ECMWF reports and wiki pages (ECMWF, 2024)

. These operational guidelines specify that, to issue a formal flood notification, EFAS first applies strict criteria known as the Condition of Access (CoA). Under CoA, a notification is only issued if the catchment area is at least 1,000 km² and lies within a partner region (O'Regan, 2024b). Furthermore, the potential flood event must be forecasted to occur between 2 and 7 days from the forecast issue time. For example, for a forecast issued on 3rd May 2025 at 00 UTC, the event must start between 5th May 2025 00 UTC and 10th May 2025 00 UTC (O'Regan, 2024b).

While the CoA established the administrative boundaries for notifications, the underlying methodology for generating notifications has evolved over the years. In the previous version (EFAS v4), a notification was issued when three consecutive forecasts showed that ≥ 30% of ensemble members (from ECMWF-ENS or COSMO-LEPS) exceeded the Q_5 threshold, provided that at least one deterministic forecast (ECMWF-HRES or DWD) also exceeded it (Smith et al., 2016). The persistence criteria, meaning the number of consecutive forecasts with positive predictions within a rolling window, were applied to prevent false notifications caused by erratic model behaviour. To upgrade the notification criteria to the current operational system, various aggregation methods were evaluated during the development phase (Smith et al., 2016). In addition to EFAS v4 and the selected current operational system, two alternative aggregation approaches were tested. A simple model average, which assigns equal weights to all providers regardless of the forecast type, and a member-proportional approach, where weights were assigned based on the number of ensemble members in each model (O'Regan, 2024b). Additionally, the exceedance threshold range (5 to 95% with 2.5% increments) and the persistence configuration (1/1, 2/4, 2/2, 3/4, 3/3) were tested to identify the optimal sensitivity for the new system (O'Regan, 2024b).

Ultimately, the system evolved to the current operational version EFAS v5.2 in August 2024. This method was selected because it had the highest F-score with a beta coefficient of 0.8, effectively giving more importance to minimizing the number of false alarms that could undermine the trust in the system (Casado-Rodríguez et al., 2025; O'Regan, 2024b). Therefore, Fβ, with a coefficient of 0.8, is included in the model analysis in this study. The current operational system applies a skill-weighted approach to combine forecasts from different model providers (O'Regan, 2024b). The weights are based on the skill

of each provider model's probabilistic accuracy and horizon. This accuracy is quantified using the Brier Score (BS), which sums the squared differences between the observed outcome ($P_{obs,t}$) and the predicted probability ($P_{pred,t}$) over time steps T (O'Regan, 2024b) (Equation 10).

$$BS = \frac{1}{T}\sum_{t=1}^{T}\left(P_{obs,t} - P_{pred,t}\right)^2 \quad (10)$$

A lower BS indicates a higher predictive skill. An inverse-power transformation with an exponent of -7 is applied to convert these raw scores into operational weights (*w*) for each numerical weather prediction provider (*nwp*) at a specific horizon (*lt*) (Equation 11). This high exponent strongly favours models with better performance (O'Regan, 2024b).

$$w_{nwp,lt} = \frac{BS_{nwp,lt}^{-7}}{\sum_{i=1}^{4} BS_{i,lt}^{-7}} \quad (11)$$

Once the weights are calculated and normalized across the available providers at each forecast horizon, the final decision to issue a notification relies on a weighted ensemble approach (O'Regan, 2024b). A binary value is assigned to each model's forecast, 1 if discharge exceeds flood thresholds, 0 otherwise. The indicators are then multiplied by the corresponding model weights and summed to produce an overall exceedance probability. If this final aggregated value is ≥ 0.5 (corresponding to a 50% exceedance probability), the system issues a formal flood notification (O'Regan, 2024b). The approximate contribution weights of each provider are detailed in Table 6, which was extracted manually from **Figure C1** (Appendix C). These will be later used to reproduce EFAS analysis in this study.

Table 6

Overview of the approximated scores of each forecast provider by forecast horizon in the EFAS v5.2 notification system.

Provider	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
DWD-HRES	0.06	0.05	0.05	0.04	0.03	0.02
ECMWF-HRES	0.12	0.11	0.10	0.10	0.12	0.10
COSMO-LEPS	0.20	0.19	0.15	0.15	0.00*	0.00*
ECMWF-ENS**	0.62	0.65	0.70	0.81	0.85	0.88

** It is 0 because the COSMO-LEPS has a maximum forecast horizon of 5 days (120 hours), so it does not contribute beyond Day 5.*

*** ECMWF-ENS also includes ECMWF-CON in the score division*

If the criteria are met, a notification is issued to the relevant national partners for local action and to the Emergency Response Coordination Centre (ERCC) (EFAS, n.d.). This ensures that, while local authorities manage the immediate event, the ERCC can simultaneously improve preparedness for

potential aid requests. For transnational river basins, notifications are also shared with national authorities that might not yet be affected by the forecasted flood (EFAS, 2025).

6.2. Developing a Surrogate Model of EFAS's Decision Logic

6.2.1. Return Period

To evaluate the performance of both the EFAS and supervised ML models, a discharge threshold was defined to distinguish flood from non-flood events. In line with the statistical approach used in EFAS, return-period thresholds were derived using the Gumbel distribution of annual maximum observed discharges. Observed daily discharge data from 1990 to 2024 were collected from all 4 gauging stations. **Table 7** reports the resulting discharge thresholds (*QT*) for return periods (*T*) of 1.5, 2, 5, and 20 years.

Table 7

Overview of the calculated return period thresholds

Station	$Q_{1.5}$ (m³/s)	Q_2 (m³/s)	Q_5 (m³/s)	Q_{20} (m³/s)
Lobith	4923	5675	7526	9928
Megen Dorp	1204	1316	1589	1945
Venlo	1245	1393	1758	2231
St Peter Noord	1375	1538	1939	2459

The discharge thresholds represent the water flow magnitude expected to be reached or exceeded, on average, once every *T* years. Because flood events are statistically rare, the frequency with which observed discharge exceeded each threshold in the study period is limited. This is shown in **Figure 7**, which plots the daily mean flow (m³/s) against the estimated return-period thresholds. The shaded grey highlights a period in which EFAS data are unavailable, spanning from late 2023 to spring 2024. This data gap coincides with a rainy season and is hydrologically active with significant flow fluctuations, which unfortunately could not be used in the analysis.

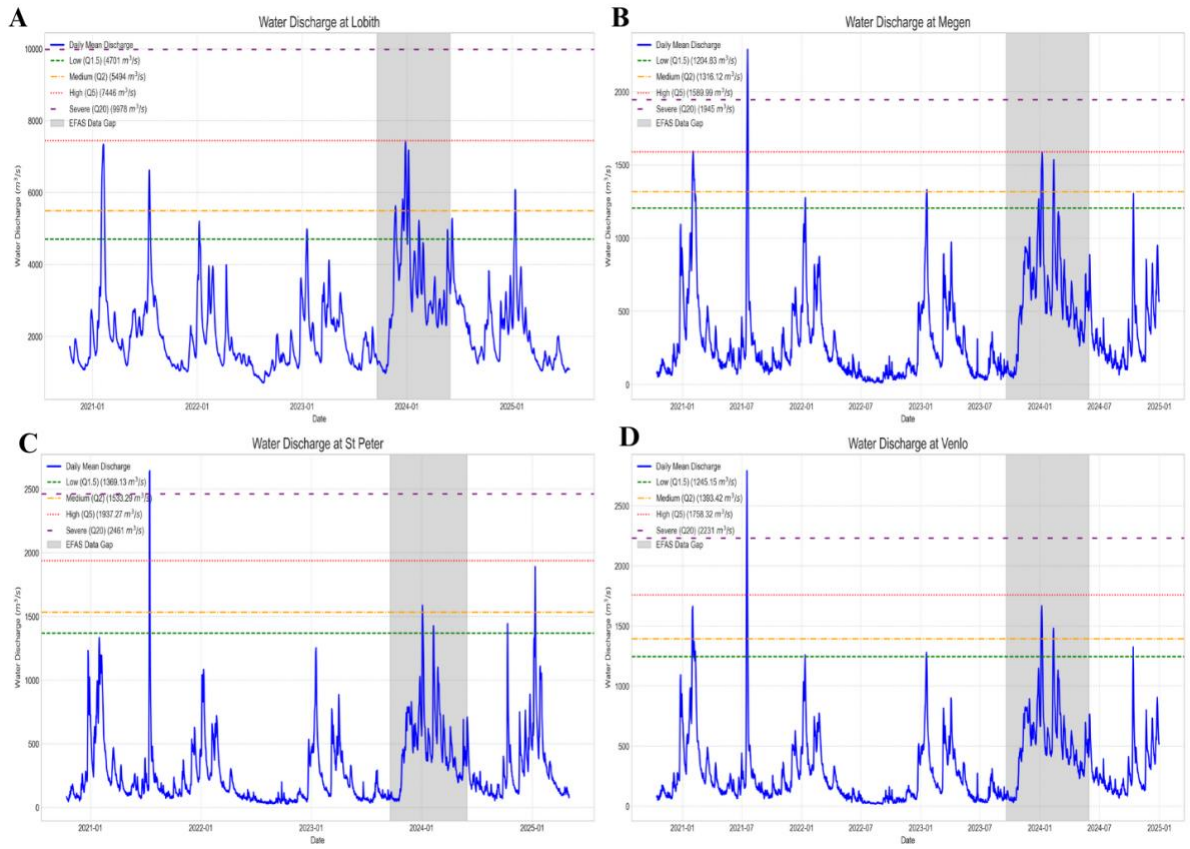

Figure 7. *Observed hydrographs relative to flood risk thresholds from October 2020 to April 2025. Coloured dashed lines indicate the $Q_{1.5}$ (green), Q_2 (orange), Q_5 (red) and Q_{20} (purple) return periods. The grey shaded region indicates the data gap in the EFAS archive. A is Lobith, B is Megen Dorp, C is St Peter, and D is Venlo.*

To quantify the data scarcity issue, **Table 8** shows the number of times the observed discharge exceeded each threshold at each station. Exceedances are reported both as counts per forecast issuance (12-hour, corresponding to the EFAS forecast cycle) and as counts of unique calendar days on which at least one exceedance occurred.

Table 8

Overview of observed discharge intervals exceeding each return -period threshold at the four gauging stations. These exceedances are based solely on the observed discharge records from October 2020 to April 2025.

Threshold	Events above the threshold (12-hour)				Events above the threshold (day)			
	Lobith	Megen	Venlo	St Peter	Lobith	Megen	Venlo	St Peter
Low ($Q_{1.5}$)	61	50	32	15	29	24	16	7
Medium (Q_2)	33	28	14	9	17	14	7	4
High (Q_5)	0	10	6	5	0	5	3	2
Severe (Q_{20})	0	6	4	2	0	3	2	2

As shown in **Table 8**, although Lobith shows the greatest number of exceedances for lower thresholds, it does not have exceedances above Q5 recorded. More generally, exceedances of Q_5 are scarce across all stations, resulting in a dataset that it is insufficient to train a supervised ML model. For this reason, although EFAS issues operational flood notifications using the Q_5 threshold, this study adopts the $Q_{1.5}$ threshold to ensure that there are enough events for model training and evaluation. Even with the adoption of the lower $Q_{1.5}$ threshold, the target class remains rare. Across the entire dataset of almost 10,400 samples, only 167 instances are classified as flood events (an imbalance ratio of ≈1.6%) The implication and limitations of the decision of using a low threshold are discussed in Section 7.

6.2.2. EFAS Performance

Following the overview of EFAS operations in the literature review (Section 6.1.) and using the calculated discharge thresholds $Q_{1.5}$ for each river gauge station, this section shows the performance of the surrogate of the current EFAS notification system to answer the second SQ: *"How accurate has the EFAS's flood notifications system been in recent years, based on historical forecast and observation data?"*. The results cover the period between June 2024 and April 2025 and are based on the aggregated forecasts and observations from the four selected river locations. The performance is assessed over forecast horizons ranging from 2 to 7 days, with results reported at a daily resolution.

Figure 8 shows the evolution of detection performance across forecast horizons. The bar graph displays the counts of TP, FN, and FP, with TN omitted due to the overwhelming class imbalance and to keep the readability of the figure. In the short-range (day 2 to day 3), the system has a positive detection rate. TP consistently exceeds FN, meaning it identifies the majority of approaching flood events. FP remains moderate, suggesting a reasonable trade-off between sensitivity and precision. However, a shift occurs at the 4th day horizon. The number of missed events (12 FN) surpasses the number of detected events (10 TP) for the first time. Beyond this point, the system's sensitivity decreases. While the number of false alarms remains relatively stable, decreasing slightly at longer horizons, the count of missed floods increases steadily. By the 7th day horizon, the FN (19) are nearly five times higher than TP (4). Therefore, while EFAS is reliable in the short term, its ability to detect flood signals decreases beyond the 4th day horizon, leading to a high rate of missed events.

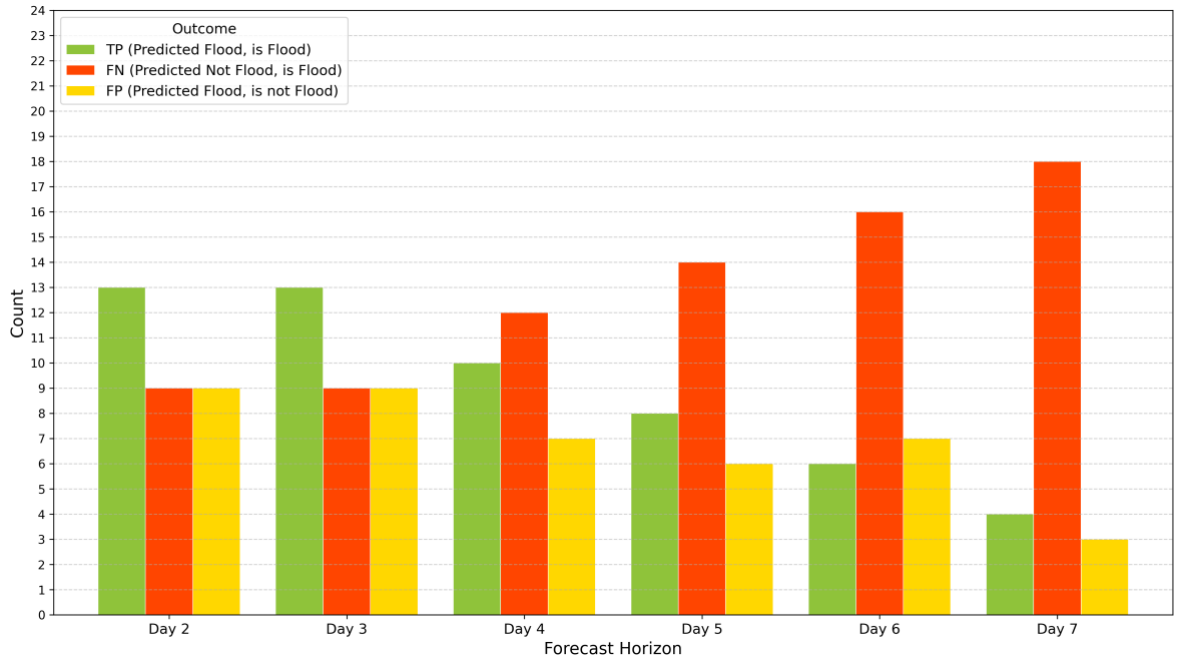

Figure 8. *Distribution of flood prediction outcomes (TP, FN, and FP) across forecast horizon.*

This trend is quantified by the performance metrics shown in the heatmap (**Figure 9**). Precision remains stable across the entire forecast period, fluctuating between 0.591 and 0.462. This indicates that the reliability of the issue alerts does not degrade with the increase of horizon, even at day 7 horizon, more than half of the warnings generated correspond to actual flood events (0.571). This stability is driven by the concurrent reduction in FP observed in **Figure 8**. As the system becomes less sensitive, it issues fewer alerts overall, hence preventing a shift in the ratio of correct to incorrect warnings. In contrast, Recall shows a steep decline with increasing forecast horizon, from 0.591 on the 2nd and 3rd days to just 0.182 at day 7. This sharp trajectory confirms that the system's performance degradation is driven almost exclusively by EFAS's inability to detect flood signals. Therefore, the F1-score also declines, as it shows the balance between precision and recall. F0.8-score, which weighs precision higher than recall, maintains a higher baseline value (0.311 on the 7th day) compared to the F1-score (0.276). This confirms that the system's performance is not due to an increase in issuing more false alarms, but rather a reduced ability to detect floods.

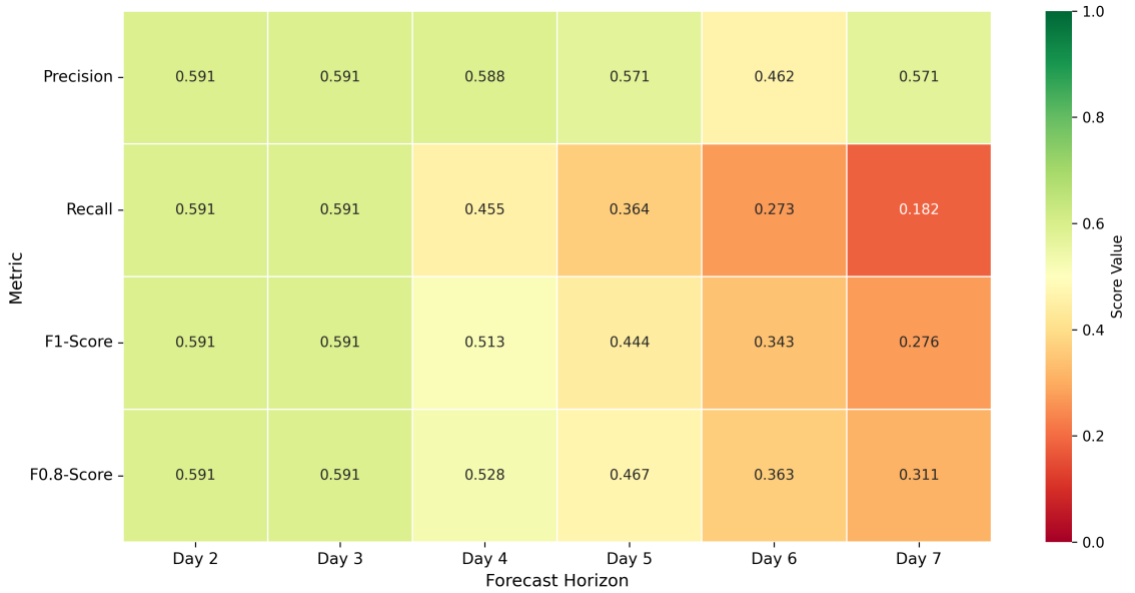

Figure 9. *Heatmap displaying the performance metrics (precision, recall, F1-score and F0.8-score) of the ML model across different forecast horizons. Green means higher performance scores, while yellow/red indicates lower scores.*

Finally, the precision-recall curves presented in **Figure 10** provide further support for these findings. Each curve represents the trade-off between recall (x-axis) and precision (y-axis) across all possible decision thresholds at a given horizon. The dashed baseline is the performance of a No-Skill classifier (0.011), which is the expected performance of a random classifier that cannot distinguish between positive and negative. It means it would guess based on the number of positives in the dataset, which in this case is 1%. Overall, as the horizon increases, the PRAUC decreases. In the short- to medium-range (day 2 to day 5), the system shows moderate detection skill, maintaining values between 0.6 and 0.5. This means that for the horizons up to the 5th day, the model is able to identify a meaningful number of flood events without sacrificing precision. Meanwhile, performance deteriorates at longer horizons (day 6 and day 7). The curves are compressed toward the bottom-left, and the PRAUC values decline to 0.412 and 0.39, which means it is impossible to improve detection rates without accepting a disproportionately high number of false alarms.

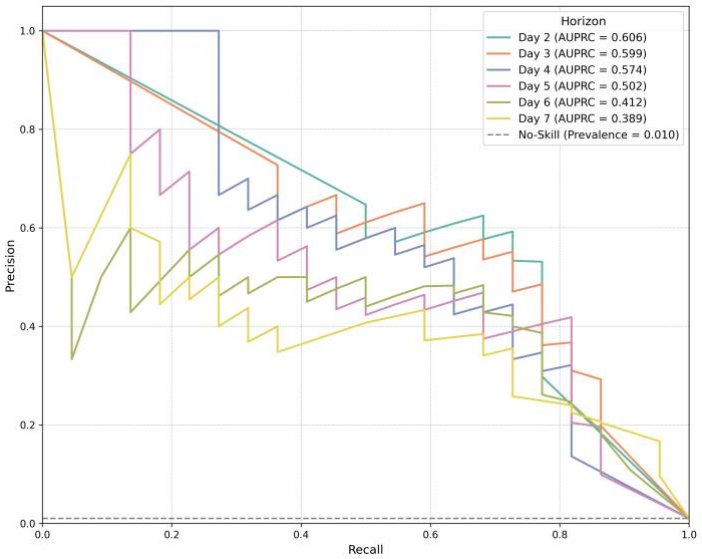

Figure 10. *Precision-Recall curves for the EFAS model across all forecast horizons. The dashed grey line represents the No-Skill baseline, indicating the expected performance of a random classifier.*

6.3. Supervised ML Model

This section evaluates the performance of the supervised ML model, addressing the research question "*How accurately can a supervised machine learning model predict flood notifications and how does its performance compare to the current operational system?"*. As with the EFAS analysis in Section 6.2, the supervised ML model's test evaluation covers the period from June 2024 to April 2025 and uses the aggregated forecasts and observations from the four selected river locations.

6.3.1. Training, Validation, and Test Loss Results

A model was trained for each prediction horizon. The stability of the optimisation process was assessed by analysing the learning curves obtained from the time-series cross-validation. Across all horizons, the learning curves show a rapid decrease in loss during the initial epochs, followed by a gradual stabilisation. After convergence, training and validation losses remain relatively stable and closely aligned, indicating consistent optimisation behaviour across folds. The corresponding training and validation loss curves for each fold and horizon are provided in Figure D1 in Appendix D.

Final model performance was evaluated on a strictly held-out test set that was not used during training or cross-validation. **Table 9** summarises the test Loss values for all forecast horizons. The lowest test losses are observed at shorter horizons, with values of 0.0241 at day 2 horizon, 0.0327 at day 3 horizon, 0.0253 at day 4 horizon, and 0.0340 at day 5 horizon. At longer horizons, an increase was observed, reaching 0.0444 for day 6 and 0.0552 for day 7. These loss values reflect a decline in model reliability as the forecast horizon increases.

Table 9

Overview of test loss values across forecast horizons

Horizon	Test Loss Value
Day 2	0.0241
Day 3	0.0327
Day 4	0.0253
Day 5	0.0340
Day 6	0.0444
Day 7	0.0552

6.3.2. Evaluation Metric Results

The evaluation metric uses the same EFAS assessments across the forecast horizons, enabling direct comparison. **Figure 11** shows the distribution of TP, FN and FP. For the short to medium range (day 2 to day 4), the model shows robust detection abilities. The number of FP is low, ranging from 3 at day 3 horizon to 5 at day 2, indicating a limited number of incorrect flood alerts. FN values remain between 8 and 10, showing that while the models identify more true events than it misses. A notable deterioration in performance occurs at day 5. The TP is at the lowest (10) and FN highest (12). At day 6, the model shows increased sensitivity but reduced precision, as while TP increases to 13, false alarms spike to 12. The model issues almost as many incorrect alerts as correct ones, so it is difficult to distinguish noise from signals. At the longest forecast horizon, predictive skills decline even further. The model generates more false alarms (14) than TP (12), and the number of missed events (FN) rises to 10. Overall, as the forecast horizon increases, the model's ability to detect flood events weakens.

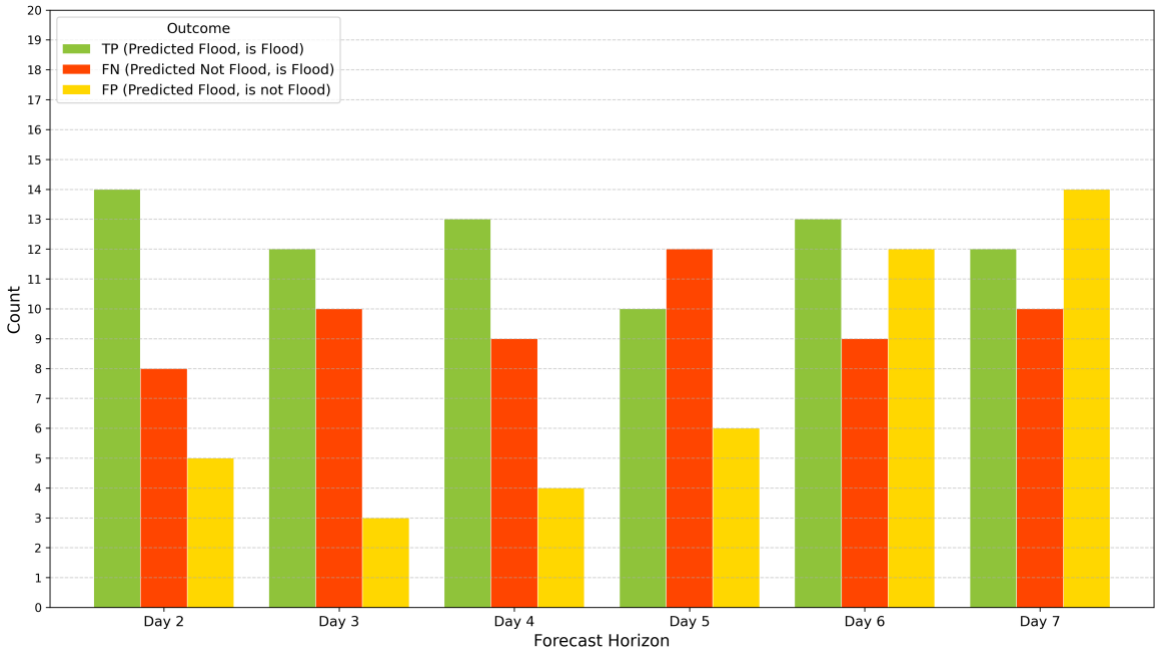

Figure 11. *Distribution of flood prediction outcomes (TP, FN, and FP) across forecast horizon.*

This trend is also confirmed by the performance metrics in **Figure 12**, which provide additional insights into the model's behaviour. In the short-to medium-range (day 2 to day 3), the model performs well. Precision is the strongest metric, starting at 0.737 (day 2), peaking at 0.8 (day 3), and staying at 0.765 (day 4). This means the model is highly reliable: most alerts it sends are real floods and have low rates of unnecessary warnings. Recall is lower, ranging from 0.636 at day 2 to 0.545 at day 3 and 0.591 at day 4. However, F1-scores are stable, between 0.649 and 0.683, showing the model is solid overall. Performance starts to drop at day 5, as precision falls to 0.625 and recall to 0.455, resulting in an F1-score of 0.526. At day 6, precision decreases to 0.520, indicating that nearly half of the alerts are incorrect. Even though it still catches floods (recall 0.591), it does so at the expense of higher false alarm rates. By the 7th day horizon, precision drops to around 0.462 and the F0.8-score, which prioritises precision, follows the same trend as the F1-score. Since the model keeps precision higher than recall, the F0.8-score confirms that overall, it is better than F1-score in each forecast, except at day 6.

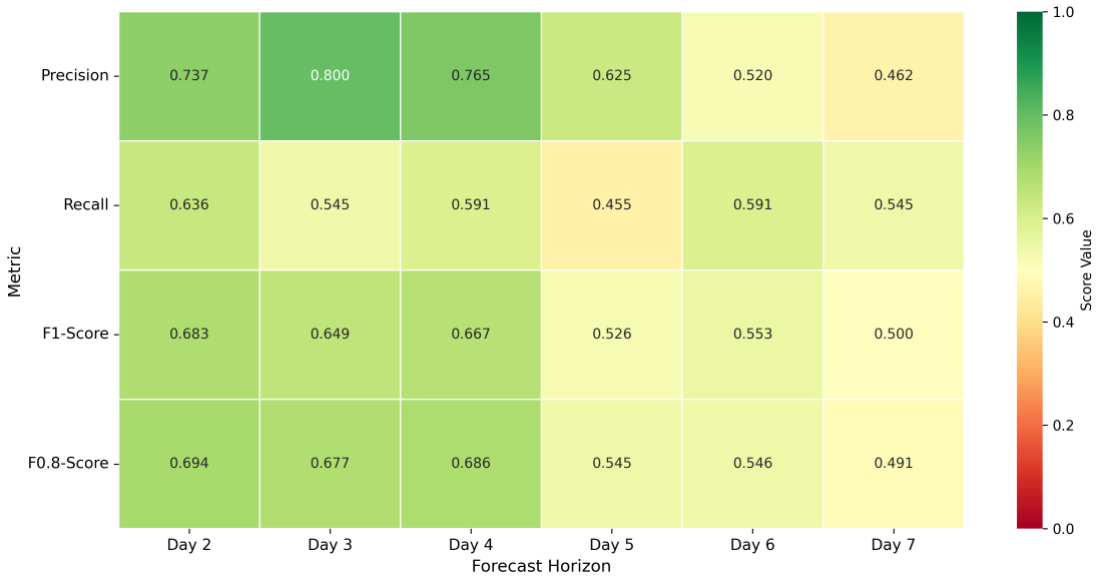

Figure 12. *Heatmap displaying the performance metrics (precision, recall, F1-score, and F0.8-score) of the ML model across different forecast horizons. Green means higher performance scores, while yellow/red indicates lower scores.*

Finally, **Figure 13** shows the trade-off between precision and recall for each forecast horizon. Generally, as the horizon increases, the model's ability to discriminate between flood and non-flood events decreases, reflected in progressively lower AUPRC values. The 2nd day horizon performs best (0.673), while the day 3 and day 4 horizons show relatively strong skill, with AUPRC at around 0.615. The AUPRC drops from 0.451 at day 5 to just 0.363 at day 7, roughly half of the skill achieved at day 2. Across all horizons, the shapes of the curves follow a similar pattern: as recall increases, precision declines sharply, meaning that capturing more flood events by lowering the decision threshold inevitably causes a higher proportion of false alarms.

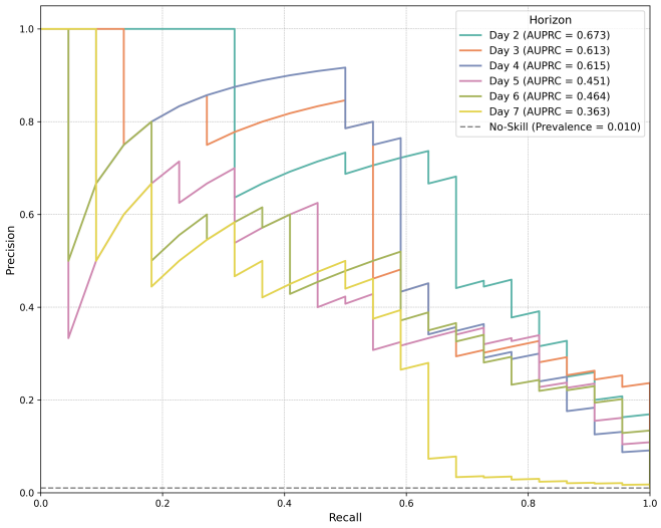

Figure 13. *Precision-Recall curves for the ML model across all forecast horizons. The dashed grey line represents the No-Skill baseline, indicating the expected performance of a random classifier.*

6.3.3 Comparison of the Models

This section compares the performance of the supervised ML model against the current EFAS operational notification approach across forecast horizons of day 2 to day 7. **Figure 14** shows the absolute difference between the two systems (Δ = ML-EFAS) for event outcomes (TP, FN, and FP) at each forecast horizon. Improvements are defined as positive values for TP, together with negative values of FN and FP.

On day 2, the ML model detects 1 additional flood event than EFAS while simultaneously reducing the number of false alarms by 4. At day 3, the ML model detects one more event than EFAS (-1 TN) and has 6 fewer false alarms. The ML model performs best in the medium range on day 4, detecting more flood events (+3 TP) while issuing fewer false alerts (-3 FP). At day 5, it maintained a detection advantage (+2 TP) without generating any additional false alarms compared to EFAS. This is the range in which the ML approach contributes the most. At extended horizons (day 6 and day 7), performance changes and the ML model become more sensitive. While it correctly identifies significantly more floods (+7 to +8 TP) and thereby reduces missed detections, it also leads to a spike in false alarms (+5 to +11 FP). This pattern confirms that, at longer horizons, the model becomes excessively sensitive, capturing signals that EFAS misses while generating higher noise.

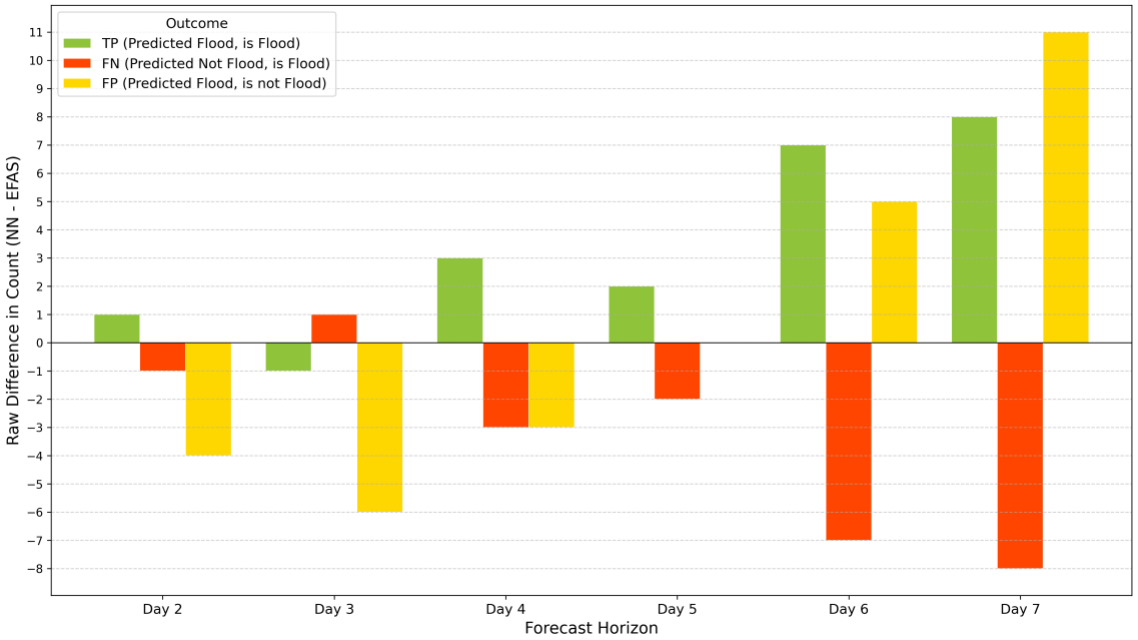

Figure 14. *Comparative analysis showing the absolute count difference between the supervised ML model and EFAS.*

Figure 15 shows the percentage improvement of the supervised ML model relative to the EFAS baseline. Positive values mean performance gains, while negative values indicate deterioration. In terms of precision, the supervised ML model is stronger across the short-to-medium range (day 2 to day 4), meaning it is more effective than EFAS at limiting false alarms in this time window. The largest improvement occurs at day 3 (+35.4%). However, this advance decreases at day 7, where precision declines by 19.2%, consistent with the increased noise observed in previous results. Recall shows the opposite trend. While the ML model is more conservative at day 3, showing a 7.7 drop in detection relative to EFAS, its sensitivity increases with horizon. Beyond day 5, recall improved dramatically, reaching +116.7% at day 6 and +200% at day 7. The trend confirms that the ML model captured a larger share of actual floods at longer horizons than EFAS. Finally, when balancing the trade-offs, both the F1-score and F0.8-score show consistent improvements across all forecast horizons. Unlike previous results, there are no negative aggregate scores. The F1-score ranges from +9.8% at day 3 to +81.2% at day 7.

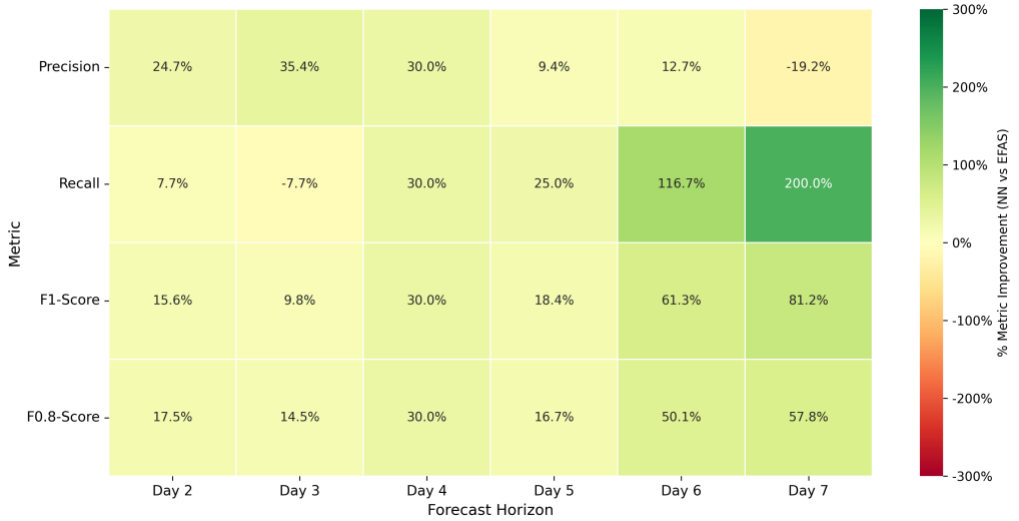

Figure 15. *Heatmap showing the relative percentage improvement of the ML model over EFAS metrics ($\frac{ML-EFAS}{EFAS}$). Green cells indicate horizons where the ML model outperforms the baseline.*

Finally, **Figure 16** illustrates the absolute difference in AUPRC between the ML model and the EFAS baseline (Δ = ML – EFAS). At day 2, the ML shows a clear advantage. The performance gain is substantial (+0.067). In the window between day 3 and day 4 hours, the predictive skill of both systems fluctuates. The performance increased slightly at day 3 (+0.014), followed by a gain at day 4 (+0.041). However, this advantage reverses at day 5, where the ML model shows its largest drop in performance relative to EFAS (-0.051). At longer horizons, the model recovers with strong performance at day 6 (+0.052), significantly outperforming EFAS, only to drop again at day 7 (-0.026). Overall, this shows the differences in the models' trend in stability across various horizons. Because AUPRC evaluates performance across all possible decision thresholds, at the specific classification threshold used for the confusion-matrix evaluation, the model can be more precise than the EFAS system.

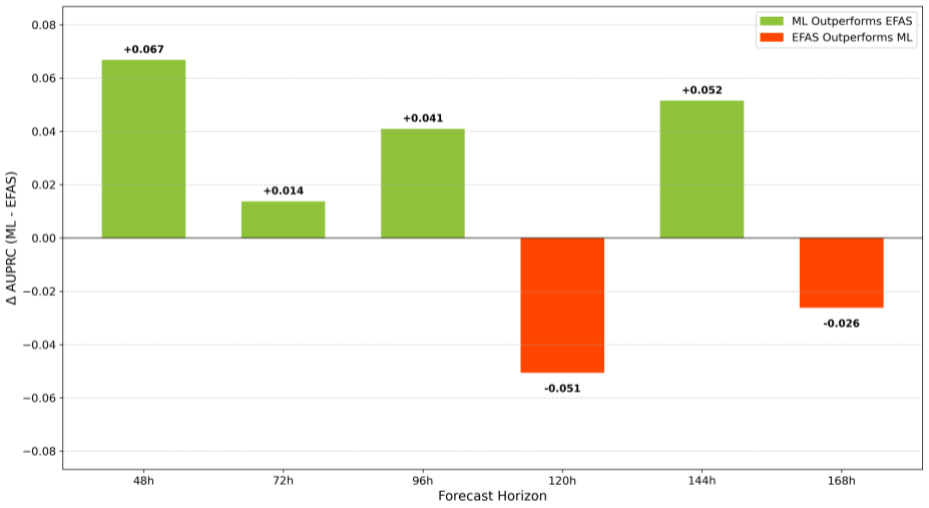

Figure 16. *Difference in AUCPR between the supervised model and EFAS. Green bars indicate horizons where the ML model has a better trade-off between precision and recall, while red bars indicate EFAS has a better trade-off.*

7. Discussion

7.1. Use of ML in Forecasting

The European Commission is increasingly launching initiatives that focus on the development of digital replicas of the Earth system, as part of the Green Deal and Digital Strategy (European Union, 2025). Within this context, recent developments in ECMWF and CESM suggest an operational shift towards hybrid forecasting frameworks that integrate physically based modelling with ML approaches. In 2024, Copernicus stated that AI algorithms can be used to help analyse Earth Observation datasets. The frequency of natural disasters will increase due to climate change and create even larger datasets (Copernicus, 2024), and a major advantage of ML is that it can be used to implicitly learn from large amounts of data (Reichstein et al., 2025). Meanwhile, ECMWF is developing digital twins, including the Weather-Induced Extremes and Climate Change Adaptation, to support risk management by improving the representation of climate change and extreme weather events (DestinE, 2025). In 2024, Anemoi was launched, a collaborative framework created by ECMWF and other services in Europe, to build an open-source ecosystem to train, test and use ML weather forecasting models (Lentze, 2025). Since September 2025, the Artificial Intelligence Forecasting System (AIFS) has also become operational within EFAS, combining numerical weather prediction with AI-driven components (Snell, 2025; Zsoter et al., 2025). From the first evaluations, AIFS delivers higher forecast skill (Zsoter et al., 2025), higher speed and efficiency, using up to 1000 times less energy than traditional physics-based models (Snell, 2025). All these developments point to a growing interest in using ML techniques as complementary tools to enhance operational forecasting abilities.

In this context, this study addressed a complementary challenge, whether a supervised ML model could improve the flood-notification component of EFAS, by enhancing the accuracy when issuing an alert. The study's results suggest that the supervised ML has the potential to enhance the notification performance, particularly at short to medium horizons, highlighting ML's potential as a complementary decision-support tool. The ML model shows its strongest advantages at horizons between 2 and 4 days, a window that aligns closely with the EFAS operational upgrade that favoured the configuration performing best between 2 and 5.5 days (Casado-Rodríguez et al., 2025). Within this range, the ML-based approach reduced FP, indicating increasing reliability, and almost always lowered FN counts compared to the surrogate EFAS baseline. Reducing missed flood events is particularly important from a risk-management and emergency response perspective, as undetected events can lead to severe societal and economic impacts. Previous evaluations of EFAS have shown that timely flood warnings can provide enough time for preparedness and response measures, which not only save lives but has returns in investments of roughly 400 times higher than the operation costs (Pappenberger et al., 2015). However, the selection of the current operational EFAS notification system was based on using the

F0.8-score as evaluation metric (O'Regan, 2024b). It prioritises precision, meaning a preference for limiting false alarms, to ensure that issued alarms are credible to users. Even under this precision-focused criterion, the ML-based approach achieves a higher F0.8-score across all forecast horizons, due to a reduction in FP in the short to medium horizon window, while at longer horizons (day 5 to day 7) due to a higher detection capability (recall) than EFAS, despite a rise in noise. Finally, the higher AUPRC achieved by the ML model in the range between 2 and 4 days, shows improved separation between flood and non-flood cases, with the model assigning higher probabilities to true flood events. Operationally, it enhances the robustness of the notification decisions.

7.2 Challenges in Flood Prediction

While the results suggest potential improvements, the challenges related to the nature of the dataset and problem structure need to be addressed. The characteristics of the data in this study posed significant challenges, influencing modelling decisions. Rare event detection is a challenging classification problem, as it is difficult to predict when the events of interest are only a tiny portion of the entire dataset (Hadi et al., 2024; He & Cheng, 2021). Floods, depending on their return period, may occur once every several years, which results in a highly imbalanced dataset. With an imbalanced dataset, the learning algorithms are biased in favour of the dominating class (He & Cheng, 2021). Therefore, ML approaches require sufficient representation of both classes to learn meaningful discriminative patterns.

Common approaches for handling imbalanced data include resampling strategies such as random undersampling or oversampling (Moniz et al., 2017). Because the dataset used here is already limited, undersampling the majority class would remove valuable information about the data distribution and weaken the model's ability to generalise to unseen datasets (Alkhawaldeh et al., 2023). Standard random oversampling strategies were also evaluated (Appendix E). Random oversampling of the minority class and point-wise resampling strategies resulted in reduced performance, with higher FN and FP despite stable loss convergence. These outcomes highlight the need to avoid treating data points as singular observations, but rather as event flood blocks. For this reason, a variation of time-series bootstrap is applied (Härdle et al., 2003). The adopted event-based oversampling strategy, in which consecutive positive events are grouped and randomly oversampled with replacement, addresses imbalance while preserving the temporal structure of each flood event. The hold-out test results in Section 6.3.1 and the validation and training loss curves in Appendix D, confirm that this method limits the risk of overfitting inherent in oversampling. The convergence of training and validation losses across all horizons, indicates stable generalization behaviour to unseen data, implying that the model learns hydrological relevant patterns rather than memorising training samples.

7.3 Differences between EFAS and ML Model

Differences in input structure between EFAS and the ML model must also be acknowledged. The weight distribution in Section 6.1. shows that EFAS relies heavily on ECMWF-ENS forecasts, with ensemble members receiving increasing weight across the full forecast horizon. This method is able to preserve the full range of the ensemble spread information of 51 members. In contrast, the supervised ML model only incorporates ECMWF-ENS as an input feature from day 5. Additionally, instead of using all ensemble members individually, the model uses the ensemble median to reduce input dimensionality and noise. This method supports a more stable learning with the small sample size available in this study. However, while this simplification reduces model's complexity and overfitting risk, it sacrifices detailed ensemble information.

In addition to differences in forecast-providers' inputs, the supervised ML model in this study integrates observed discharge values from the 12 hours preceding each forecast issue time. This feature introduces hydrological near-real-time observation in the input dataset. Operationally, it would mean that the model has access to real measurements closest to the point at which a decision must be made, improving its ability to distinguish early stages of rising discharge across forecast horizons between day 2 and day 4 (Konold et al., 2025). Although EFAS only issues notifications strictly from the forecast products, the CEMS Hydrological Data Collection Centre (HDCC) has access to in-situ real-time and historical discharge and/or water levels, covering 52% of European water basins (Padilla et al., 2025). The direct inclusion of additional context can improve flood-alert predictions but introduces structural dependencies with gauging coverage. In regions with data scarcity across temporal and spatial scales, it may hinder model robustness between different catchments (Nie et al., 2025). It reflects a broad trend in flood forecasting and water management, where methodological improvements in models must be supported by robust observational infrastructure, datasets, and institutional coordination. This aligns with the current CESM-HDCC plans (Padilla et al., 2025).

Another difference from the EFAS operational system concerns the threshold definitions. In EFAS, return-period thresholds are from simulated discharge, while this study's return periods are from observed discharge data. Simulation-based thresholds are higher than observation-based and are affected by different error characteristics. The choice to use observed discharge-based thresholds was made to ensure internal consistency between the surrogate EFAS implementation, the supervised model and the observational dataset used for evaluation.

7.4. Limitations & Future Work

Despite the promising results achieved by the ML model, this study is subject to several limitations. First, ML models are often perceived as *"black boxes"* because they are difficult to interpret due to their complexity (Jain et al., 2018; Kumar et al., 2025; Nie et al., 2025). The inherent limited physical

interpretability of the model and the lack of transparency about which features are responsible for the outcome, can reduce understanding of the system and hinder trust among stakeholders (Nie et al., 2025). Another limitation is that the study was applied only in the Netherlands, therefore, the transferability of the results to other regions remains to be tested.

However, the study's main limitation concerns the limited dataset size, approximately 10,000 samples for training, validation, and testing, which is considered relatively small for NNs. Given the complexity of the problem, a NN was an appropriate starting point, but such models typically require a substantially larger number of samples to learn complex relationships. Sufficiency of a dataset is important both in size and representativeness (Hatamian et al., 2025). Small datasets are more sensitive to changes in initialization values, training procedures or small differences in data samples used during training (Arbel et al., 2023; Hatamian et al., 2025).

The data constraint also influenced the definition of the binary ground truth. The operational EFAS, floods notifications are based on the Q_5 return-period threshold. Within the temporal and spatial scope of this study, Q_5 -level events are extremely rare, making it infeasible to train or meaningfully evaluate the supervised ML model. Therefore, to ensure that the ML model had a minimum number of positive samples, a lower threshold ($Q_{1.5}$) was used. As the objective of this study was to evaluate and improve the decision logic that translates forecasts into binary notifications, rather than to issue real operational flood alerts, this choice was necessary for methodological feasibility. Nevertheless, it also means that the binary classification is targeting high-flow events, rather than the extreme events (Smith et al., 2016), which should be considered when interpreting the results and their applicability.

Future works could address these limitations by expanding the dataset either temporally, by allowing the flood events to accumulate over the next years, or spatially, by adding additional catchments and extending the analysis to a European scale. A larger and more diverse dataset could enable the use of thresholds that are consistent with EFAS's operational definition. Having also a larger volume of data for the ML model, it would also enable the possibility of more advanced deep learning architectures, such as Long Short-Term Memory (LSTM) networks. These models are designed to capture complex temporal dependencies but need substantial data to train effectively.

8. Conclusion

The study investigated whether a supervised ML model could improve the accuracy of the decision-making component of EFAS. The analysis focused on the notification system, where EFAS converts hydrological forecasts into a binary alert that indicates whether a forecasted event is a flood. To answer the main research question, three SQ were formulated. First, SQ1 aimed at providing an overview how EFAS currently issues flood notifications. The analysis showed that EFAS applies a threshold-based decision rule. The current operational system aggregates forecasts using a skill-weighted ensemble, where forecasts from different providers are assigned weights based on Brier Scores across all horizons. An alert is issued when the weighted probability of exceeding the Q_5 return period threshold reaches at least 0.5. Following, SQ2 evaluated the accuracy of the current EFAS notification system by using historical forecasts and observational data. The study confirmed that EFAS performs reliably at short horizons, with increasing uncertainty at longer horizons. Most errors relate to missing events. Finally, SQ3 analysed the supervised ML model and compared its performance with EFAS. The supervised ML approach achieved clear improvements at short-to-medium horizons, reducing both FN and FP, thereby improving the ability to detect actual flood events without increasing unnecessary alerts. The model also showed consistent behaviour across the validation and test sets, meaning that it learned relevant patterns. However, performance weakened by day 5, as false alarms became more frequent. Overall, the results show that the supervised ML model can be used to enhance EFAS's flood notification accuracy within the operationally critical 2- to 4-day forecast horizon. Based on these results, it is recommended to evaluate the approach across a broader European spatial study area and assess its robustness and general applicability. Such extension would allow the ML-based decision support to be tested as a component of the EFAS notification system.

9. References

Adams, T. E., & Pagano, T. C. (2016). *Flood forecasting: a global perspective*. Elsevier/Ap, Academic Press Is An Imprint Of Elsevier.

Alfieri, L., Burek, P., Feyen, L., & Forzieri, G. (2015). Global warming increases the frequency of river floods in Europe. *Hydrology and Earth System Sciences*, *19*(5), 2247–2260. https://doi.org/10.5194/hess-19-2247-2015

Alkhawaldeh, I. M., Albalkhi, I., & Naswhan, A. J. (2023). Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, *13*(5), 373–378. https://doi.org/10.5662/wjm.v13.i5.373

Anghel, C. G. (2024). Revisiting the Use of the Gumbel Distribution: A Comprehensive Statistical Analysis Regarding Modeling Extremes and Rare Events. *Mathematics*, *12*(16), 2466–2466. https://doi.org/10.3390/math12162466

Arbel, J., Pitas, K., Vladimirova, M., & Fortuin, V. (2023). *A Primer on Bayesian Neural Networks: Review and Debates*. ArXiv.org. https://arxiv.org/abs/2309.16314

ATLAS.ti Scientific Software Development GmbH. (2023). ATLAS.ti Mac (version 23.2.1) [Qualitative data analysis software]. https://atlasti.com

Bentivoglio, R., Isufi, E., Sebastian Nicolaas Jonkman, & Taormina, R. (2021). *Deep Learning Methods for Flood Mapping: A Review of Existing Applications and Future Research Directions*. https://doi.org/10.5194/hess-2021-614

Bianchi, V. (2021). Cars are submerged in floodwaters after the Meuse River broke its banks during heavy flooding in Liege, Belgium, Thursday, July 15, 2021. Heavy rainfall is causing flooding in several provinces in Belgium with rain expected to last until Friday. [Photograph]. In *Phys.org*. https://phys.org/news/2021-07-europe-toll-tops.html

Biesbroek, R., Schmidt, D., Alexander, P., Børsheim, K., Carnicer, J., Georgopoulou, E., Haasnoot, M., Le Cozannet, G., Lionello, P., Lipka, O., Möllmann, C., Muccione, V., Mustonen, T., Piepenburg, D., Pörtner, H.-O., Roberts, D., Tignor, M., Poloczanska, E., Mintenbeck, K., & Alegría, A. (2022). Lorraine Whitmarsh (UK) Contributing Authors. *Ana Mijic*, 1817–1927. https://doi.org/10.1017/9781009325844.015

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. https://people.sabanciuniv.edu/berrin/cs512/lectures/Book-Bishop-Neural%20Networks%20for%20Pattern%20Recognition.pdf#page=131.22

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., & Gorbachova, L. (2019). Changing climate both increases and decreases European river floods. *Nature*, *573*(7772), 108–111. https://doi.org/10.1038/s41586-019-1495-6

Bomers, A., Schielen, R. M. J., & Hulscher, S. J. M. H. (2019). Decreasing uncertainty in flood frequency analyses by including historic flood events in an efficient bootstrap approach. *Natural Hazards and Earth System Sciences*, *19*(8), 1895–1908. https://doi.org/10.5194/nhess-19-1895-2019

Byaruhanga, N., Kibirige, D., Gokool, S., & Mkhonta, G. (2024). Evolution of Flood Prediction and Forecasting Models for Flood Early Warning Systems: A Scoping Review. *Water*, *16*(13), 1763–1763. https://doi.org/10.3390/w16131763

Casado-Rodríguez, J., Carton de Wiart, C., Grimaldi, S., Zsoter, E., Baugh, C., Bosshard, N., Mikuličková, M., Pechlivanidis, I., Prudhomme, C., & Salamon, P. (2025). Optimizing Warnings for Hydrologic Ensemble Prediction Systems for Improved Decision-Making. *Journal of Hydrometeorology*, *26*(6), 675–689. https://doi.org/10.1175/jhm-d-24-0054.1

Chamola, V., Hassija, V., Gupta, S., Goyal, A., Guizani, M., & Sikdar, B. (2020). Disaster and Pandemic Management Using Machine Learning: A Survey. *IEEE Internet of Things Journal*, 1–1. https://doi.org/10.1109/jiot.2020.3044966

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1). https://doi.org/10.1186/s13040-023-00322-4

Cloke, H., & Pappenberger, F. (2008). *574 Operational flood forecasting: a review of ensemble techniques*. https://www.ecmwf.int/sites/default/files/elibrary/2008/8742-operational-flood-forecasting-review-ensemble-techniques.pdf?

Copernicus. (2024, May 9). *OBSERVER: Revealing hidden land patterns with AI and Copernicus | Copernicus*. Copernicus.eu. https://www.copernicus.eu/en/news/news/observer-revealing-hidden-land-patterns-ai-and-copernicus

Copernicus Climate Change Service (C3S), & World Meteorological Organization (WMO). (2025). *European State of the Climate 2024*. climate.copernicus.eu/ESOTC/2024, doi.org/10.24381/14j9-s541

Copernicus Emergency Management System (CEMS). (2019). *EFAS-IS* [Map]. https://european-flood.emergency.copernicus.eu/efas_frontend/#/home

CRED. (2020, October 12). *The human cost of disasters: an overview of the last 20 years (2000-2019)*. Undrr.org. https://www.undrr.org/quick/50922

De Roo, A. P. J., Wesseling, C. G., & Van Deursen, W. P. A. (2000). Physically based river basin modelling within a GIS: the LISFLOOD model. *Hydrological Processes*, *14*(11-12), 1981–1992. https://doi.org/10.1002/1099-1085(20000815/30)14:11/12%3C1981::aid-hyp49%3E3.0.co;2-f

DestinE. (2025, March 5). *About*. Destination Earth. https://destine.ecmwf.int/about/

Deepa, S., Siddalingappa, R., Kalpana, P., Loveline Zeema, J., Vinay, M., Jayapriya, J., & Priya Stella Mary, I. (2025). A Comparative Analysis of L1, L2, and L1L2 Regularization Techniques in Neural Networks for Image Classification. *International Journal of Engineering Trends and Technology, 73*(10). https://ijettjournal.org/archive/ijett-v73i10p108

Doocy, S., Daniels, A., Murray, S., & Kirsch, T. D. (2013). The Human Impact of Floods: a Historical Review of Events 1980-2009 and Systematic Literature Review. *PLoS Currents*, *5*(1). https://doi.org/10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a

Doyle, E. E. H., Johnston, D. M., Smith, R., & Paton, D. (2019). Communicating model uncertainty for natural hazards: A qualitative systematic thematic review. *International Journal of Disaster Risk Reduction*, *33*, 449–476. https://doi.org/10.1016/j.ijdrr.2018.10.023

ECMWF. (2020). *Ensemble weather forecasting - Fact Sheet*. https://www.ecmwf.int/sites/default/files/medialibrary/2017-03/ecmwf-fact-sheet-ensemble-forecasting.pdf

ECMWF. (2024). *CEMS-Flood - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki*. Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/CEMS-Flood

EEA. (2025). *Climate change*. Europa.eu. https://discomap.eea.europa.eu/climatechange/?page=Floods&views=Recent-events--

EFAS. (n.d.). *EUROPEAN FLOOD AWARENESS SYSTEM (EFAS) CONDITIONS OF ACCESS between the EFAS Dissemination Centre*. https://european-flood.emergency.copernicus.eu/sites/default/files/documents/EFAS_Condition_of_Access_new_final.pdf

EFAS. (2024). *European Flood Awareness System EFAS Bulletin*. https://european-flood.emergency.copernicus.eu/sites/default/files/bulletins-documents/2024/EFAS_Bimonthly_Bulletin_Jun_Jul2024.pdf

EFAS. (2025). *CEMS-Floods React*. Efas.eu. https://stage.efas.eu/react/faq

European Union. (2025). *Destination Earth | Shaping Europe's digital future*. Digital-Strategy.ec.europa.eu. https://digital-strategy.ec.europa.eu/en/policies/destination-earth

Fakhruzi, I. (2018). An artificial neural network with bagging to address imbalance datasets on clinical prediction. *2018 International Conference on Information and Communications Technology (ICOIACT)*, 895–898. https://doi.org/10.1109/icoiact.2018.8350824

Fares, A., Alkhodre, A. B., Adnan, Muhammad Sher Rama, Alzahrani, B., & Muhammad Shoaib Siddiqui. (2023). Flood Prediction using Hydrologic and ML-based Modeling: A Systematic Review. *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, *14*(11). https://doi.org/10.14569/ijacsa.2023.0141155

Farhadi, Z., Bevrani, H., & Feizi-Derakhshi, M.-R. (2022). *Combining regularization and dropout techniques for deep convolutional neural network*. 335–339. https://doi.org/10.1109/GEC55014.2022.9986657

Fergus, P., & Chalmers, C. (2022). Performance Evaluation Metrics. *Computational Intelligence Methods and Applications*, 115–138. https://doi.org/10.1007/978-3-031-04420-5_5

Feyen, L., Ciscar, J., Gosling, S., Ibarreta, D., & Soria, A. (2020). Climate change impacts and adaptation in Europe: JRC PESETA IV final report. *Publications Office of the European Union*. https://doi.org/10.2760/171121

Ha, K., Cho, S., & MacLachlan, D. (2005). Response models based on bagging neural networks. *Journal of Interactive Marketing*, *19*(1), 17–30. https://doi.org/10.1002/dir.20028

Hadi, A., Lariyah Mohd Sidek, Salih, A., Hidayah Basri, Saad Sh. Sammen, Norlida Mohd Dom, Zaharifudin Muhamad Ali, & Ali Najah Ahmed. (2024). Machine learning techniques for flood forecasting. *Journal of Hydroinformatics*. https://doi.org/10.2166/hydro.2024.208

Härdle, W., Horowitz, J., & Kreiss, J. (2003). Bootstrap Methods for Time Series. *International Statistical Review*, *71*(2), 435–459. https://doi.org/10.1111/j.1751-5823.2003.tb00485.x

Hatamian, A., Levine, L., Ehsani, O. H., & Sarrafzadeh, M. (2025). *Exploring the Impact of Dataset Statistical Effect Size on Model Performance and Data Sample Size Sufficiency*. ArXiv.org. https://arxiv.org/abs/2501.02673

He, J., & Cheng, M. X. (2021). Weighting Methods for Rare Event Identification From Imbalanced Datasets. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.715320

Henley, J., & Jones, S. (2024, October 31). *Spain floods death toll passes 150 as country begins three days of mourning*. The Guardian; The Guardian. https://www.theguardian.com/world/2024/oct/31/spain-floods-valencia-death-toll-three-days-mourning

Hidayat Jati, M. I., Suroso, & Santoso, P. B. (2019). Prediction of flood areas using the logistic regression method (case study of the provinces Banten, DKI Jakarta, and West Java). *Journal of Physics: Conference Series*, *1367*(1), 012087. https://doi.org/10.1088/1742-6596/1367/1/012087

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01-11. https://doi.org/10.5121/ijdkp.2015.5201

Hyndman, R. J., & Athanasopoulos, G. (2018). 3.4 Evaluating forecast accuracy | Forecasting: Principles and Practice. In *Forecasting: principles and practice,.* 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Ilemobayo, J. A., Durodola, O., Alade, O., Awotunde, O. J., Olanrewaju, A. T., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., Odezuligbo, I. E., & Edu, O. E. (2024). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. *Journal of Engineering Research and Reports*, *26*(6), 388–395. https://doi.org/10.9734/jerr/2024/v26i61188

Islam, M., Chen, G., & Jin, S. (2019). An Overview of Neural Network. *American Journal of Neural Networks and Applications*, *5*(1), 7. https://doi.org/10.11648/j.ajnna.20190501.12

Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., Kumar, S., Agarwal, S. P., & Dimri, A. P. (2018). A Brief review of flood forecasting techniques and their applications. *International Journal of River Basin Management*, *16*(3), 329–344. https://doi.org/10.1080/15715124.2017.1411920

Joint Research Center (JRC), & Copernicus Emergency Management Service (CEMS). (2019). *Iver discharge and related forecasted data from the European Flood Awareness System. Early Warning Data Store (EWDS)*. http://doi.org/10.24381/cds.9f696a7a

Jones, S. (2024). *"Catastrophe of epic proportions": eight drown in Europe amid heavy floods*. The Guardian; The Guardian. https://www.theguardian.com/world/2024/sep/15/catastrophe-of-epic-proportions-six-drown-in-europe-amid-heavy-floods-storm-boris-poland-austria-slovakia-hungary

Jonsson, F., & Rydén, J. (2017). 5 STATISTICAL STUDIES OF THE BETA GUMBEL DISTRIBUTION: ESTIMATION OF EXTREME LEVELS OF PRECIPITATION. *Statistica Applicata -Italian Journal of Applied Statistics*, *29*(1). https://www.diva-portal.org/smash/get/diva2:1128358/FULLTEXT01.pdf

Khan, S. M., Shafi, I., Butt, W. H., Diez, I. de la T., Flores, M. A. L., Galán, J. C., & Ashraf, I. (2023). A Systematic Review of Disaster Management Systems: Approaches, Challenges, and Future Directions. *Land*, *12*(8), 1514. https://doi.org/10.3390/land12081514

Kingma, D. P., & Ba, J. (2014, December 22). *Adam: A Method for Stochastic Optimization*. ArXiv.org. https://arxiv.org/abs/1412.6980

Konold, O., Feigl, M., Podest, P., Klingler, C., & Schulz, K. (2025). BiasCast: Learning and adjusting real time biases from meteorological forecasts to enhance runoff predictions. *EGUsphere*. https://doi.org/10.5194/egusphere-2025-4978

Kumar, V., Sharma, K. V., Mangukiya, N. K., Tiwari, D. K., Ramkar, P. V., & Rathnayake, U. (2025). Machine learning applications in flood forecasting and predictions, challenges, and way-out in the perspective of changing environment. *AIMS Environmental Science*, *12*(1), 72–105. https://doi.org/10.3934/environsci.2025004

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., Mach, K., Muir-Wood, R., Brakenridge, G. R., Kron, W., Benito, G., Honda, Y., Takahashi, K., & Sherstyukov, B. (2013). Flood Risk and Climate change: Global and Regional Perspectives. *Hydrological Sciences Journal*, *59*(1), 1–28. https://doi.org/10.1080/02626667.2013.857411

LeClerc, J., & Joslyn, S. (2015). The Cry Wolf Effect and Weather-Related Decision Making. *Risk Analysis*, *35*(3), 385–395. https://doi.org/10.1111/risa.12336

Lentze, G. (2024, October 14). *Anemoi: a new framework for weather forecasting based on machine learning*. ECMWF.https://www.ecmwf.int/en/about/media-centre/news/2024/anemoi-new-framework-weather-forecasting-based-machine-learning

Lentze, G. (2025, February 25). *ECMWF's AI forecasts become operational*. ECMWF. https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational

Li, D., Fang, Z. N., & Bedient, P. B. (2021). Chapter 6 - Flood early warning systems under changing climate and extreme events. In A. Fares (Ed.), *Climate Change and Extreme Events* (pp. 83–103). Elsevier. https://doi.org/10.1016/B978-0-12-822700-8.00002-0

Li, X., Liang, X., Wang, X., Wang, R., Shu, L., & Xu, W. (2023). Deep reinforcement learning for optimal rescue path planning in uncertain and complex urban pluvial flood scenarios. *Applied Soft Computing*, *144*, 110543–110543. https://doi.org/10.1016/j.asoc.2023.110543

Machiwal, D., & Jha, M. K. (2012). Introduction. In *n: Hydrologic Time Series Analysis: Theory and Practice.* (pp. 1–12). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1861-6_1

Maskell, K. (2023, June 13). *The rise of machine learning in weather forecasting*. ECMWF. https://www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting

Matthews, G., Barnard, C., Cloke, H., Dance, S. L., Jurlina, T., Mazzetti, C., & Prudhomme, C. (2022). Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System. *Hydrology and Earth System Sciences*, *26*(11), 2939–2968. https://doi.org/10.5194/hess-26-2939-2022

Mazzetti, C., Carton de Wiart, C., Gomes, G., Russo, C., Decremer, D., Ramos, A., Grimaldi, S., Disperati, J., Ziese, M., Schweim, C., Sanchez Garcia, R., Jacobson, T., Salamon, P., Prudhomme, C. (2023): River discharge and related historical data from the European Flood Awareness System, v5.0. European Commission, Joint Research Centre (JRC). DOI:10.24381/cds.e3458969

Mishra, A., Mukherjee, S., Merz, B., Singh, V. P., Wright, D. B., Villarini, G., Paul, S., Kumar, D. N., Khedun, C. P., Niyogi, D., Schumann, G., & Stedinger, J. R. (2022). An Overview of Flood Concepts, Challenges, and Future Directions. *Journal of Hydrologic Engineering*, *27*(6). https://doi.org/10.1061/(asce)he.1943-5584.0002164

Moniz, N., Branco, P., & Torgo, L. (2017). Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics*, *3*(3), 161–181. https://doi.org/10.1007/s41060-017-0044-3

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water*, *10*(11), 1536. https://doi.org/10.3390/w10111536

Muñoz, P., Orellana-Alvear, J., Bendix, J., Feyen, J., & Célleri, R. (2021). Flood Early Warning Systems Using Machine Learning Techniques: The Case of the Tomebamba Catchment at the Southern Andes of Ecuador. *Hydrology*, *8*(4), 183. https://doi.org/10.3390/hydrology8040183

Muñoz, P., Orellana-Alvear, J., Willems, P., & Célleri, R. (2018). Flash-Flood Forecasting in an Andean Mountain Catchment—Development of a Step-Wise Methodology Based on the Random Forest Algorithm. *Water*, *10*(11), 1519. https://doi.org/10.3390/w10111519

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. *Lecture Notes in Networks and Systems*, *724*, 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., & Reich, O. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, *26*(15), 4013–4032. https://doi.org/10.5194/hess-26-4013-2022

Newaz, A., Hassan, S., & Haq, F. S. (2022, August 24). *An Empirical Analysis of the Efficacy of Different Sampling Techniques for Imbalanced Classification*. ArXiv.org. https://doi.org/10.48550/arXiv.2208.11852

Nie, Y., Yu, K. H., Wang, Y., & Liu, P. (2025). Applications of machine learning and deep learning in hydrology from a bibliometric perspective: a comprehensive review. *Discover Artificial Intelligence*, *5*(1). https://doi.org/10.1007/s44163-025-00471-x

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Neuralnetworksanddeeplearning.com; Determination Press. http://neuralnetworksanddeeplearning.com/chap1.html#sigmoid_neurons

O'Regan, K. (2023). *EFAS hydrological model performance - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki*. Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/EFAS+hydrological+model+performance

O'Regan, K. (2024a). *EWDS API - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki*. Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/EWDS+API

O'Regan, K. (2024b, October 30). *EFAS v5.2 - updates - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki*. Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/EFAS+v5.2+-+updates

O'Regan, K. (2024c, November 27). *EFAS - Known Issues - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki*. Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/EFAS+-+Known+Issues

O'Regan, K. (2025). *EFAS Meteorological forecasts - Copernicus Emergency Management Service - CEMS - ECMWF Confluence Wiki* (E. Zsoter, Ed.). Ecmwf.int. https://confluence.ecmwf.int/display/CEMS/EFAS+Meteorological+forecasts

Owens, R. G., & Hewson, T. D. (2018). ECMWF Forecast User Guide. *ECMWF*. https://doi.org/10.21957/m1cs7h

Padilla, G., Sánchez, G., Prestigiacomo, J., Molina, J., Arroyo, M., Márquez, S., & Grimaldi, J. (2025). The CEMS Hydrological Data Collection Centre Annual Report 2023. *Publications Office of the European Union, Luxembourg,*. https://doi.org/10.2760/0123431

Panarin, R. (2024, January 15). *[Basic Data Augmentation Method applied to Time Series]*. Custom Software Development Company; Mad Devs Group LTD. https://maddevs.io/writeups/basic-data-augmentation-method-applied-to-time-series/

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., & Thielen, J. (2015). The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, *51*, 278–291. https://doi.org/10.1016/j.envsci.2015.04.016

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J. T., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Luca Antiga, Alban Desmaison, Kopf, A., Yang, E. S., DeVito, Z., Raison, M., Tejani, A., Sasank Chilamkurthy, Steiner, B., Fang, L., & Bai, J. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1912.01703

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

PyTorch. (2024). *BCEWithLogitsLoss — PyTorch 2.7 documentation*. Pytorch.org. https://docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

Pozo, J. T., Salamon, P., Burek, P., Florian Pappenberger, Eklund, C. A., E. Sprokkereef, M. Hazlinger, Garcia, M. P., & R. Garcia-Sanchez. (2015). Medium Range Flood Forecasting Example EFAS. *Springer EBooks*, 1–17. https://doi.org/10.1007/978-3-642-40457-3_51-1

Reichstein, M., Benson, V., Blunk, J., Camps-Valls, G., Creutzig, F., Fearnley, C. J., Han, B., Kornhuber, K., Rahaman, N., Schölkopf, B., Tárraga, J. M., Vinuesa, R., Dall, K., Denzler, J., Frank, D., Martini, G., Nganga, N., Maddix, D. C., & Weldemariam, K. (2025). Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, *16*(1). https://doi.org/10.1038/s41467-025-57640-w

Rijkswaterstaat. (2025). *Rijkswaterstaat Waterinfo*. Waterinfo.rws.nl. https://waterinfo.rws.nl/nav/bulkdownload/menu

Rjkswaterstaat. (2024). *Waterstanden en waterafvoer bij Lobith*. Rijkswaterstaat.nl. https://www.rijkswaterstaat.nl/water/waterdata-en-waterberichtgeving/waterdata/lobith-waterstanden-en-afvoeren

Šakić Trogrlić, R., van den Homberg, M., Budimir, M., McQuistan, C., Sneddon, A., & Golding, B. (2022). Early Warning Systems and Their Role in Disaster Risk Reduction. *Towards the "Perfect" Weather Warning*, 11–46. https://doi.org/10.1007/978-3-030-98989-7_2

Scikit-learn. (n.d.). *3.1. Cross-validation: evaluating estimator performance*. Scikit-Learn. https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split

Scikit-learn. (2019). *sklearn.preprocessing.OneHotEncoder — scikit-learn 0.22 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., Rusticucci, M., Semenov, V., Alexander, L. V., Allen, S., Benito, G., & Cavazos, T. (2019). Changes in Climate Extremes and their Impacts on the Natural Physical Environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, 109–230. https://doi.org/10.1017/cbo9781139177245.006

Sheng, V. S., & Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. *National Conference on Artificial Intelligence*, 476–481.

Shrestha, N. (2024). *AN ACTIVE SET METHOD FOR A NEURAL NETWORK (ASM-NN)*. https://ufdcimages.uflib.ufl.edu/UF/E0/06/12/82/00001/Shrestha_N.pdf#page=14.35

Shyalika, C., Wickramarachchi, R., & Sheth, A. P. (2024). A Comprehensive Survey on Rare Event Prediction. *ACM Computing Surveys*. https://doi.org/10.1145/3699955

Siadati, S. (2021). Supervised Learning Models. In *Machine Learning: Theory and Practice* (pp. 23–39). European Organization for Nuclear Research. https://doi.org/10.5281/zenodo.16999765

Smith, P., Florian Pappenberger, Fredrik Wetterhall, Pozo, del, Krzeminski, B., Salamon, P., D. Muraro, Kalas, M., & Baugh, C. M. (2016). On the Operational Implementation of the European Flood Awareness System (EFAS). *Elsevier EBooks*, 313–348. https://doi.org/10.1016/b978-0-12-801884-2.00011-6

Snell, C. (2025). *Simplifying AI for weather forecasting with the European Weather Cloud*. ECMWF. https://www.ecmwf.int/en/about/media-centre/science-blog/2025/simplifying-ai-weather-forecasting-european-weather-cloud

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210x.13140

Tanim, A. H., McRae, C. B., Tavakol-Davani, H., & Goharian, E. (2022). Flood Detection in Urban Areas Using Satellite Imagery and Machine Learning. *Water*, *14*(7), 1140. https://doi.org/10.3390/w14071140

Thelen, T., Anarde, K., Dietrich, J. C., & Hino, M. (2024). Wind and rain compound with tides to cause frequent and unexpected coastal floods. *Water Research*, *266*, 122339–122339. https://doi.org/10.1016/j.watres.2024.122339

Tian, Y., Shu, M., & Jia, Q. (2021). Artificial Neural Network. *Encyclopedia of Mathematical Geosciences*, 1–4. https://doi.org/10.1007/978-3-030-26050-7_44-1

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*(2), 123–137. https://doi.org/10.1080/15715124.2008.9635342

UNDRR. (2016, December 1). *Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction*. Www.undrr.org. https://www.undrr.org/quick/11605

United Nations Office for Disaster Risk Reduction (UNDRR). (2017). *The Sendai Framework Terminology on Disaster Risk Reduction. "Structural and non-structural measures."* Undrr.org. https://www.undrr.org/terminology/structural-and-non-structural-measures.

Wallemacq, P., & House, R. (2018, October 10). *Economic losses, poverty & disasters: 1998-2017*. Undrr.org. https://www.undrr.org/quick/11678

WMO. (2022, March 21). *Early Warning systems must protect everyone within five years*. World Meteorological Organization. https://wmo.int/news/media-centre/early-warning-systems-must-protect-everyone-within-five-years

Wu, J., Liu, H., Wei, G., Song, T., Zhang, C., & Zhou, H. (2019). Flash Flood Forecasting Using Support Vector Regression Model in a Small Mountainous Catchment. *Water*, *11*(7), 1327. https://doi.org/10.3390/w11071327

Ye, Y., Li, Y., Ouyang, R., Zhang, Z., Tang, Y., & Bai, S. (2023). Improving machine learning based phase and hardness prediction of high-entropy alloys by using Gaussian noise augmented data. *Computational Materials Science*, *223*, 112140–112140. https://doi.org/10.1016/j.commatsci.2023.112140

Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., & Tseng, H.-W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of Hydrology*, *552*, 92–104. https://doi.org/10.1016/j.jhydrol.2017.06.020

Zsoter, E., Chevallier, M., Prudhomme, C., & O'regan, K. (2025). *AI takes CEMS flood forecasting into a new era*. ECMWF. https://www.ecmwf.int/en/newsletter/185/news/ai-takes-cems-flood-forecasting-new-era

Appendix

Appendix A: Data Preparation

This appendix outlines how EFAS raw files were converted into a structured dataset suitable for ML. The primary input data is the gridded forecast output from EFAS, which includes deterministic and ensemble forecasts from ECMWF (HRES, ENS, CON), DWD-HRES and COSMO-LEPS. Since the raw files downloaded from the platform in NetCDF4 and GRIB2, a single extraction was used with the *"xarray"* and *"cfgrib"* libraries to standardize variable names, normalize time dimensions and convert the multidimensional data into simple CSV files. Because the data was arranged in grids, the closest grid point was matched to each of the four target stations. To maintain spatial accuracy, an Euclidean distance check was applied and any point further than the 0.1º tolerance was excluded. Furthermore, a difference in grid setup was found for the Megen and Lobith stations starting in 2024. To address this, the extraction logic was adjusted to move the target coordinates for files dated 2024 onwards, ensuring that the CSV time series remained spatially consistent despite the change. Following the extraction into CSV, the files were aggregated into unified daily records. To ensure that the ML model had consistent input, the dataset was filtered to keep only the dates that contained values from all system's providers: ECMWF-HRES, ECMWF-ENS, ECMWF-CON, DWD-HRES and where possible COSMO-LEPS (since it has a shorter horizon range). Anomalous data was also removed. The target variables were derived from observational water flow data. As observations were in local time (CET/MET), all timestamps were converted to UTC to match the forecast issue times. Finally, a positive class (1) is assigned if the observed water level at the valid forecast time exceeds the critical thresholds $Q_{1.5}$ defined for each station.

Appendix B: Hyperparameter Configuration

This appendix provides a detailed overview of the final hyperparameter configurations selected for the ML models across all the forecast horizons. These parameters were determined through and iterative empirical approach. For each forecast horizon, model configurations were evaluated using the same time-series cross-validation scheme to ensure consistency. Performance was assessed using classification metrics and validation loss behaviour.

Table B1

Overview of the final tuned values for each hyperparameter in the ML models across forecast horizons.

Hyperparameters	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Model Architecture						
Deterministic Forecasting systems	DWD, ECMWF-CON, ECWMF-HRES	DWD, ECMWF-CON, ECWMF-HRES	DWD, ECMWF-HRES	DWD, ECMWF-HRES	DWD	DWD
Ensemble Forecasting systems	COSMO-LEPS	COSMO-LEPS	COSMO-LEPS	ECMWF-ENS	ECMWF-ENS	ECMWF-ENS
12h Prior Observation	True	True	True	False	False	False
Neurons (Hidden Layer 1)	12	13	12	14	16	20
Neurons (Hidden Layer 2)	6	7	6	7	8	10
Neurons (Hidden Layer 3)	2	2	2	2	2	4
Optimization & Loss						
Decision Threshold (τ)	0.55	0.4	0.3	0.67	0.58	0.51
Positive Class Weight (α)	0.7	0.55	0.5	0.45	0.5	0.55
Learning Rate	0.01	0.01	0.01	0.01	0.01	0.01
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Batch Size	500	500	500	500	500	500
Data Augmentation						
Target Ratio (Minority Class)	0.4	0.4	0.35	0.25	0.4	0.3
Gaussian Noise (σ)	0.021	0.02	0.015	0.03	0.02	0.005
Regularization						
L2 Penalty (λ_2)	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04
L1 Penalty (λ_1)	1.0E-05	1.0E-04	1.0E-04	1.0E-04	1.0E-04	1.0E-04

The tuned hyperparameters have a clear trend as the forecast horizon increases. For short horizons (day 2 to day 4), the input feature set includes observed water flow from 12 hours prior the issue of the forecasts to improve the model performance. These periods also included the model COSMO-LEPS and a diverse set of deterministic providers. From day 5 onwards, antecedent discharge observations

are excluded, and ensemble information is provided only by ECMWF-ENS. Deterministic inputs are progressively reduced, with ECMWF-HRES included until day 5, but only DWD stays consistently until day 6 and day 7. Short ranges perform better with a compact architecture (12-6-2 neurons), while longer horizons need deeper and wider networks, reaching 20-10-4 neurons at day 7. This increased capacity is necessary to capture more complex, non-linear relationships as system uncertainty grows over time. However, the common factor across all horizons is the funnel architecture, where the number of neurons decreases in subsequent layers to force the system to compress features before the final single output.

Most optimization parameters remained stable across all horizons, with the exception of the positive class weight (α) and the decision threshold (τ). At day 2, α was higher (0.7) compared to the longer horizons. (≈ 0.55). This means that short-range forecasts required a stronger penalty on FN to maximize detection, while for the other horizons, increasing α at these horizons failed to reduce the FN but increased the FP. Meanwhile, the decision threshold showed a significant variation at medium-range horizons, particularly at day 4, where a significantly lower threshold (0.3) was found to be beneficial, suggesting that here the model outputs lower raw probabilities for flood events.

Regarding data augmentation, the target ratio for the minority class is generally around 0.4 and dropped to 0.25 at day 5 and 0.3 at day 7, where more aggressive oversampling introduced more noise. Meanwhile, the Gaussian noise changes slightly depending on the horizon. The highest noise level (σ = 0.03) was used at day 5 and lowest at day 3 and day 7 (0.01 and 0.005). Finally, the regularization hyperparameters penalties λ_2 and λ_1 remained constant at 1.0E-04, except λ_1 at day 2 , which was set to 1.0E-05.

Appendix C: Brier Scores Weights

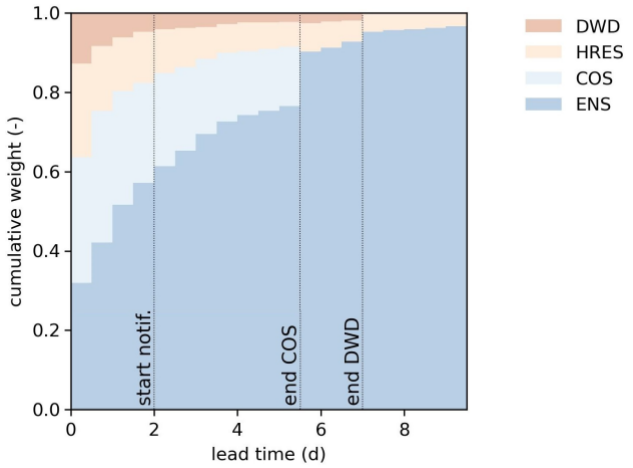

Figure C1. *Overview of the optimized notification criteria and threshold for the EFAS v5.2 upgrade across forecast horizons, adapted from the ECMWF CESM Wiki (O'Regan, 2024b).*

Appendix D: Training and Validation Losses in each Fold

This appendix presents the training and validation loss curves obtained during the time-series cross-validation procedure for all forecast horizons. **Figure D1** provides evidence of optimisation behaviour and convergence stability across validation folds. For each forecast horizon, the average of loss curves of the training loss without regularisation and the validation loss are shown of each fold. These graphs support the assessment of training stability.

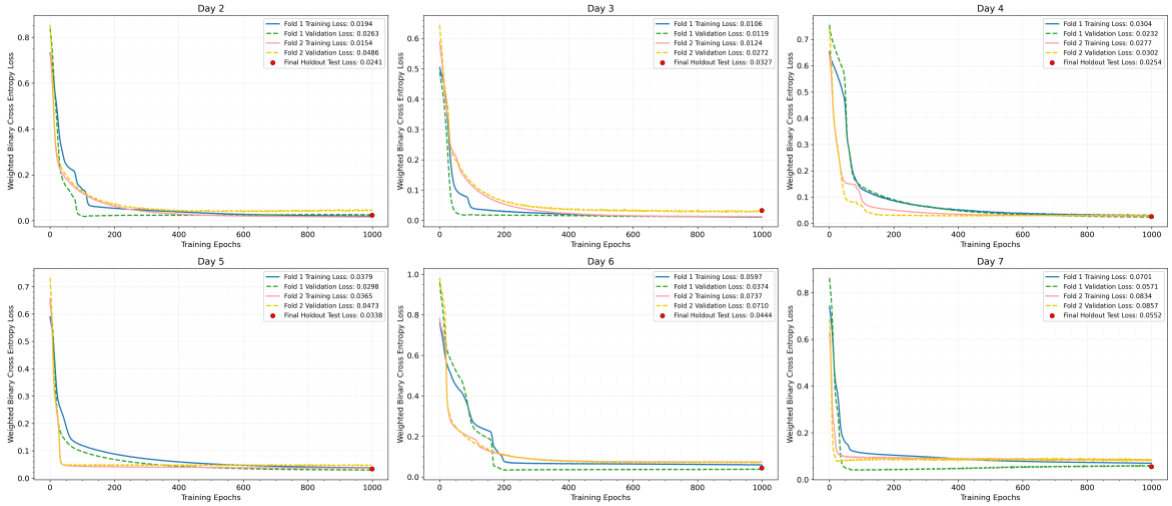

Figure D1. *Training, validation and final hold-out test loss curves for the supervised ML model across forecast horizon from day 2 to day 7. For each horizon, the average losses of training (solid line) and of validation (dashed line) and the final hold-out test (red dot).*

Appendix E: Day 2 Horizon Analysis

This section will show additional results from analysis done on day 2 horizon when looking at what data augmentation to use in the model. This horizon was selected because it is the closest to the forecast issue time, meaning the meteorological input data should contain the least amount of uncertainty compared to longer horizons. Because alternative data augmentation strategies can alter the effective class exposure during training and bias the learned decision boundary, their impact was evaluated not only through loss convergence but also through confusion matrices.

The first method that was analysed, randomly oversampled the minority class to achieve the wanted ratio. **Figure E1** shows the validation loss and training loss. Despite the high positive class weight, the model missed 14 flood events (FN) while correctly identifying only 8 (FP) and generating 9 FP

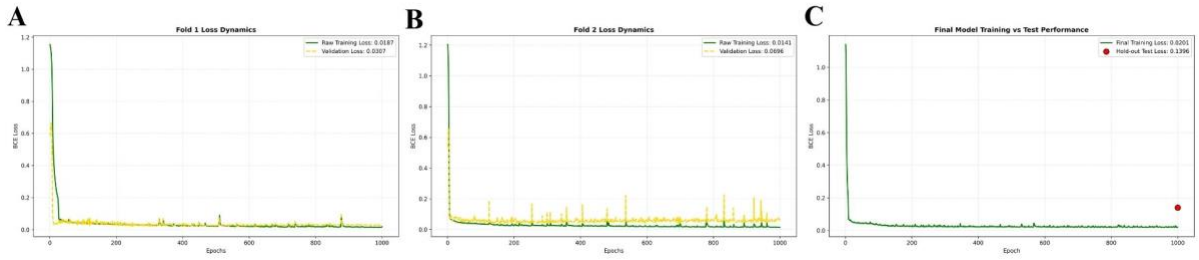

Figure E1. *Training and validation loss dynamics for day 2 forecast horizon using random oversampling. Panel A shows the first fold, B the second fold, and C the final training with hold-out test-set loss.*

A second method was analysed, which aimed at having more samples for the model to train on. This strategy first duplicates the majority class (non-floods samples) to increase the diversity of negative samples exposed to the model. Then it oversampled the minority class to match the target ratio. The FN dropped to 8 but FP rose to 10.

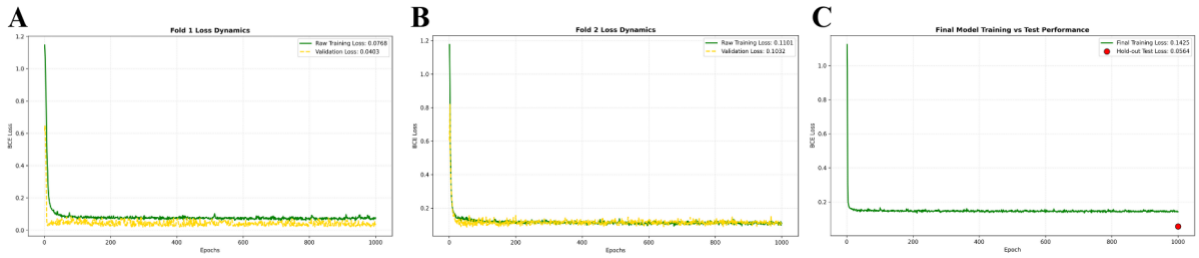

Figure E2. *Training and validation loss dynamics for day 2 forecast horizon by doubling the majority class prior to random oversampling. Panel A shows the first fold, B the second fold, and C the final training with hold-out test-set loss.*

The selected event-based oversampling approach consistently achieved lower validation loss after convergence compared to alternative methods. When combined with a more favourable trade-off between FN and FP in the validation data as well, this motivated its selection as the final augmentation strategy.

Appendix F: Additional Figures

To show the temporal behaviour of EFAS and ML-based forecasts at the event level, **Figures F1** and **F2** provide timelines for the Day 2 and Day 7 forecast horizons. These horizons were selected to represent, respectively, the short-range regime in which model skill is highest and the long-range regime in which forecast uncertainty and decision errors are the highest.

Figure F1. *Temporal comparison of flood prediction outcomes at day 2 forecast horizon across four river gauging stations. The timelines show the performance of the operational EFAS system (top row crosses) and the ML model (middle row squares) against the observed Ground Truth (bottom row vertical markers). Panels display selected high flow period for Lobith (A), Megen (B), Venlo (C), St. Peter (D). Colour indicates classification performance, with green being TP, oranges being FP and red FN.*

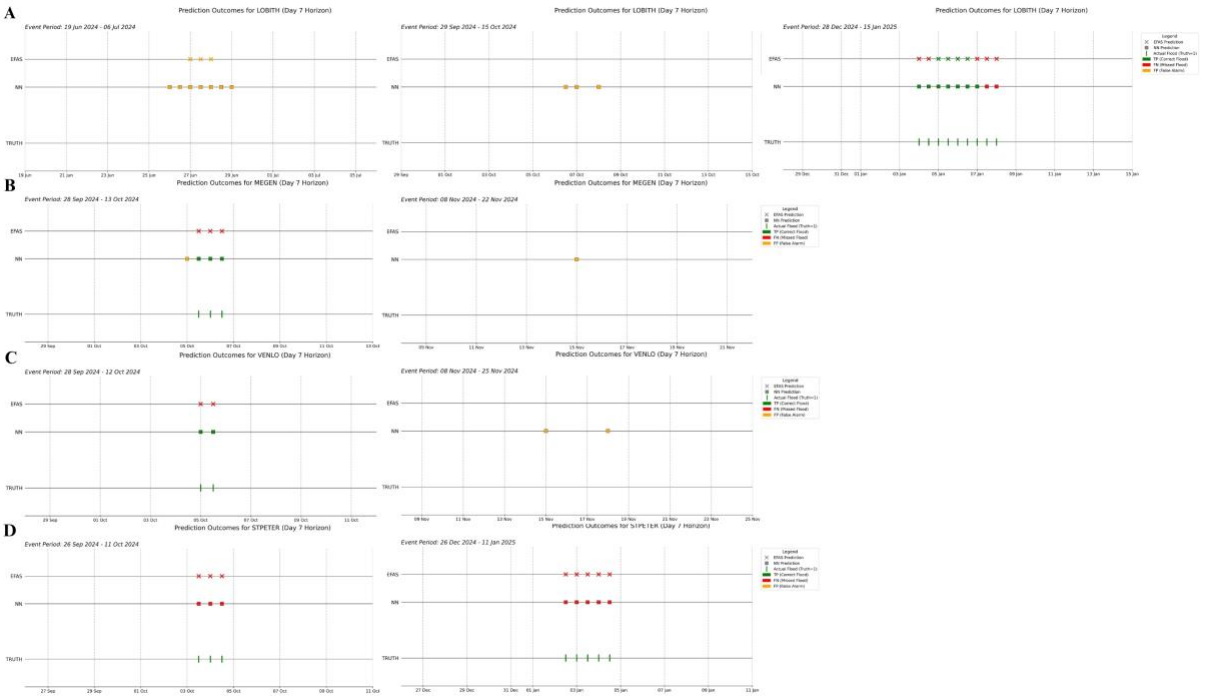

Figure F2. *Temporal comparison of flood prediction outcomes at day 7 forecast horizon across four river gauging stations. The timelines show the performance of the operational EFAS system (top row crosses) and the ML model (middle row squares) against the observed Ground Truth (bottom row vertical markers). Panels display selected high flow period for Lobith (A), Megen (B), Venlo (C), St. Peter (D). Colour indicates classification performance, with green being TP, orange being FP and red FN.*

Appendix G: Code Availability

https://git.wur.nl/barbi008/mscthesiscode