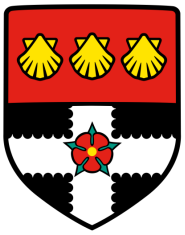

UNIVERSITY OF READING
Department of Geography & Environmental Science
School of Archaeology, Geography and Environmental Science

&

EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
Evaluation Section, Forecasts and Services Department

Outrunning flash floods: improving global medium-range forecasts for better preparedness

Fatima Maria Pillosu

Thesis submitted for the degree of Doctor of Philosophy
July 2025

DECLARATION

I, Fatima M. Pillosu, confirm that this is my work, and the use of all material from other sources has been properly and fully acknowledged.

Fatima M. Pillosu
February 27, 2026

ABSTRACT

The UN's "Early Warning for All" initiative, supported by WMO, prioritises flash floods due to their high mortality rates, widespread exposure, major economic impacts, and climate-driven exacerbation. Global medium-range predictions are essential for protecting vulnerable populations. However, despite recent advances in medium-range numerical weather prediction and hydrological modelling, developing medium-range predictions of areas at risk of flash floods over a continuous global domain remains severely constrained by computational requirements, data availability, and the inherent challenge of predicting localised extreme events.

This thesis aims to develop a first proof-of-concept of medium-range (up to day 5) predictions of areas at risk of flash floods over a continuous global domain. To achieve this goal, three interconnected research objectives are addressed using the Continental United States (CONUS) as the primary study region. The selection of CONUS leverages the Storm Event Database, which provides a comprehensive long-term record of flash flood impact reports essential for robust model development and validation.

First, a flash-flood-focused verification framework, directly comparing rainfall predictions and flash flood impact reports, is developed. This framework is used throughout the thesis, but it is first used to assess whether rainfall forecasts from global NWP models can identify areas at risk of flash floods up to medium-range lead times. The ERA5 reanalysis and ERA5 forecasts, post-processed with the ecPoint technique, show good reliability and discrimination ability up to day 5.

Second, multiple data-driven models—including random forest, gradient boosting, and neural networks—are evaluated to determine their capacity to extract predictive signals using severely imbalanced observational datasets. These models integrate hydro-meteorological variables from the ERA5 reanalysis and forecasts with flash flood impact reports to identify patterns indicative of flash flood risk. When compared to rainfall-based predictions alone, the

data-driven hydro-meteorological predictions demonstrate superior performance while maintaining computational efficiency suitable for operational deployment. Among the tested models, the XGBoost implementation of gradient boosting emerges as the best performer.

Third, systematic sensitivity analysis demonstrates that it is possible to deploy such a regionally-trained data-driven model with global hydro-meteorological forecasts to produce medium-range predictions of areas at risk of flash floods over a continuous global domain.

This research helps to establish methodologies to extract valuable predictive information from limited observational datasets and lower-resolution hydro-meteorological forecasts that could enhance preparedness and emergency management strategies, contributing to the UN's "Early Warnings for All" initiative.

ACKNOWLEDGEMENTS

I would first and foremost like to extend my deepest gratitude to my supervisors at the University of Reading, Hannah Cloke and Elisabeth Stephens, and my co-supervisor at ECMWF, Christel Prudhomme. Their unwavering support, advice and guidance throughout these eight years have been instrumental, extending beyond professional mentorship. They supported me when I decided to start building my family during my PhD and provided invaluable encouragement during moments of self-doubt whilst navigating the challenges of balancing motherhood with academic and professional commitments. I am also grateful to the Water@Reading research group for stimulating discussions across diverse subjects.

I am profoundly indebted to Florian Pappenberger and Tim Hewson, who facilitated my appointment as a graduate trainee at ECMWF nearly a decade ago and cultivated my enthusiasm for flash flood prediction, which ultimately inspired this doctoral research. I particularly acknowledge Ivan, whose lunchtime discussions have provided both cherished friendship and invaluable insights into convection. I am equally grateful to Thomas Haiden and David Richardson for their expert guidance in navigating forecast verification. I extend particular gratitude to Mariana Claire for introducing me to machine learning; this thesis would not have taken its current direction without her invaluable support and guidance. Xavi Abellan and Loriano Pagni deserve particular thanks for their assistance with the technical aspects of this thesis. They always helped me when ”the computer said no”. I finally want to extend my appreciation to my writing coach, Melanie Smith. As a non-English speaker, I had serious difficulties in structuring complex arguments in English at the beginning of my PhD. Through her assistance, I was able to develop my academic writing skills and enjoy writing as I do in my mother tongue.

I am grateful to numerous colleagues and friends at ECMWF and the University of Reading, including Sinead, Alan, Claudia V., Rebecca, Louise, Esti, Cinzia, Calum, Ervin, David, Fernando, Cihan, Axel, Francesca, Fredrik, Claudia DN., Giacomo, Giovanna, and Laura. Even if

they were not directly involved in my thesis, they all helped me to go through the most stressful moments of my PhD through a good laugh or (many) pints of cider. It is true, after all, that it takes a village!

Last but not least, I want to acknowledge the unwavering support of my family: my parents, my sister Claudia, and my brother-in-law Adrian. I would not have arrived at the end of this period without my beloved husband, Tim. I do not only owe you what I know about the clouds, radar, and synoptic maps; you have been my rock in the most challenging moments. Most profoundly, I dedicate this achievement to my daughters, Isabel and Sofia, and to the two I was unable to meet, who have been my greatest source of strength during this transformative period. Despite the demands this journey placed upon our family, your enthusiastic support never wavered.

CONTENTS

1 General introduction **1**

1.1 Research questions and objectives, and contributions to knowledge 5

1.2 Thesis structure 9

2 Literature review **11**

2.1 Can global medium-range NWP rainfall forecasts successfully identify areas at risk of flash floods? 11

2.2 Are medium-range, data-driven hydro-meteorological predictions of areas at risk of flash floods feasible using reanalysis, forecasts, and impact flash flood reports with low spatial resolution? 18

2.3 How does the coverage-density trade-off influence training data strategies to develop predictions of areas at risk of flash floods over a continuous global domain 23

3 Integrated experimental strategy **27**

3.1 Data requirements 27

3.1.1 Requirements for observational data 29

3.1.2 Requirements for hydro-meteorological reanalysis and forecast data . . . 30

3.2 Requirements for the development of data-driven models trained on severely imbalanced datasets . . . 31

3.3 Requirements for a robust verification of probabilistic predictions of areas at risk of flash floods . . . 32

4 Datasets 35

4.1 Study domain: the CONUS . . . 36

4.2 Flash flood impact reports - NOAA's Storm Event Database . . . 37

4.3 ERA5 . . . 41

4.3.1 Parameter representing the antecedent soil moisture: percentage of soil maximum saturation . . . 42

4.3.2 Parameter representing the orographic steepness: the standard deviation of the filtered sub-grid orography . . . 46

4.3.3 Parameter representing the vegetation coverage: the leaf area index (LAI) 47

4.4 ERA5-ecPoint: post-processed ERA5 rainfall with the ecPoint post-processing technique . . . 49

4.5 Case Study used throughout the thesis: Storm Ida . . . 52

5 Flash-flood-focused verification of rainfall-based predictions of areas at risk of flash floods 59

5.1 Introduction . . . 59

5.2 Development of the flash-flood-focused objective verification framework for rainfall-based predictions of areas at risk of flash floods . . . 61

5.2.1 Pre-processing of observational and forecast data . . . 61

5.2.2 Building the contingency table for probabilistic forecasts, using non-standard observations (impact reports) . . . 66

5.2.3 Verification scores . . . 68

5.2.3.1 Reliability 69
5.2.3.2 Discrimination ability 70
5.3 Results 72
5.3.1 Overall verification scores 72
5.3.2 Breakdown verification scores 72
5.4 Case study on Storm Ida 78
5.5 Discussion 79
5.6 Conclusion 81

6 Data-driven hydro-meteorological predictions of areas at risk of flash flood: from short- to medium-range lead times 83

6.1 Introduction 83
6.2 Data and methods 85
6.2.1 Feature and target variables 85
6.2.2 Development of data-driven models 85
6.2.2.1 Model architecture selection 85
6.2.2.2 Feature engineering 89
6.2.2.3 Repeated nested cross-validation for model training 90
6.2.2.4 Hyperparameter tuning 92
6.2.2.5 Loss functions 92
6.2.3 Objective verification framework 93
6.2.4 Physical interpretation of the data-driven model outputs: Shapley values . 95
6.3 Results 96
6.3.1 Model training 96

6.3.1.1 Hyperparameter importance 100
6.3.2 Verification results over reanalysis data 104
6.3.3 Verification results over forecast data 110
6.4 Physical interpretation of the data-driven model behaviour 113
6.5 Case Study: Storm Ida 116
6.6 Discussions 118
6.7 Conclusions 122

7 Towards predictions over a continuous global domain: global implementation of regionally-trained models 125

7.1 Introduction 125
7.2 Data 127
7.3 Methods 127
7.3.1 Training approaches 127
7.3.2 Training data configuration 129
7.3.3 Model configuration 129
7.3.4 Performance evaluation 129
7.4 Results 130
7.5 Case Study over the CONUS: Storm Ida 133
7.6 Case studies for regions outside the CONUS 136
7.6.1 Flash floods in Spain in October 2024 136
7.6.2 Flash floods in China in July 2021 140
7.7 Discussions 142
7.8 Conclusions 148

8 General discussions **151**

8.1 Development of a flash-flood-focused verification framework for predictions of areas at risk of flash flood against flash flood impact reports 152

8.1.1 Key insights and contributions 153

8.1.2 Limitations 154

8.1.3 Future research directions 155

8.2 Development of data-driven hydro-meteorological predictions of areas at risk of flash flood 156

8.2.1 Key insights and contributions 157

8.2.2 Limitations 159

8.2.3 Future research directions 161

8.3 Global implementation of regionally-trained data-driven models for flash flood prediction 163

8.3.1 Key insights and contributions 163

8.3.2 Limitations 165

8.3.3 Future research directions 166

9 General conclusions **169**

LIST OF FIGURES

1.1 **Global Implementation of WMO's Flash Flood Guidance System (FFGS).** The map shows in blue the countries that have adopted the FFGS as indicated in (Georgakakos et al., 2022) and the WMO's Global FFGS Status at https://experience.arcgis.com/experience/bb4357054fb5475a9a4cf688c4180454/page/FFGS-Status?views=Flash-Flood-Component%2CFlash-Flood-Component 3

1.2 **Thesis' roadmap.** Three-tier chapter framework comprising **Foundational** Chapters (Chapter 1-4: General Introduction, Literature Review, Integrated Experimental Strategy, and Datasets), **Main Analysis** Chapters (Chapter 5: Development of a flash-flood-focused verification framework; Chapter 6: Development of medium-range data-driven hydro-meteorological predictions of areas at risk of flash floods; Chapter 7, blue: Global implementation of regionally-trained data-driven models), and **Synthesis** Chapters (Chapter 8-9: General Discussions and General Conclusions). Research questions and main output(s) are indicated for each main analysis chapter. 8

2.1 **Key meteorological ingredients for flash flood-producing storms.** Panel (a) shows the key ingredient number 1, i.e., the ample and persistent moisture supply. Panel (b) shows key ingredient n.2, i.e., the uplift of the moist air (convective, frontal, orographic, and convergence uplift). Panel (c) shows the key ingredient n.2, i.e., the mechanisms for persistent rainfall over an area. . . . 13

3.1 **Thesis' integrated experimental strategy.** The upper panel of the infographic reminds the reader about the hierarchical relationship between research questions RQ1 (addressed in the Main Analysis Chapter 5, card in pink), RQ2 (Main Analysis Chapter 6, card in yellow), and RQ3 (Main Analysis Chapter 7, card in blue) and corresponding research objectives. The horizontal dashed arrows beneath each card indicate the cross-chapter information flow as introduced in Chapter 1. The lower panel of the infographic (within the solid black box) identifies the three core methodological decisions that inform the integrated experimental strategy: data source selection (Section 3.1), forecast verification strategy (Section 3.3), and data-driven model development strategy (Section 3.2). The coloured indicators on the right of each grey card identify the main analysis chapter in which each methodological decision was applied. 28

4.1 **CONUS domain.** The figure shows the orography at 1 km resolution (in shades of green and brown) and the location of the 25 most populated cities (black dots) over the CONUS. 36

4.2 **Flash flood reports in NOAA's Storm Event Database.** Panel (a) shows the point flash flood report frequencies for each grid-box within the CONUS, between 2001-2024. Four quadrants shown: North-West (NW, orange shades), North-East (NE, green), South-East (SE, blue), and South-West (SW, yellow). Shades light to dark indicate frequencies between 0–0.1%, 0.1–1%, and 1–10%. Pie chart shows the overall frequency in each quadrant and the total number of point reports (108903). Panel (b) shows the annual timeseries (1950–2024) of point flash flood reports: all flood types (grey bars), flash floods (dark red), and flash floods with latitude (lat)/longitude (lon) coordinates (red). Panel (c) shows the 2021 timeseries of point (red) and gridded (black) flash flood reports (over 24-hourly accumulation periods ending at 00 UTC). The blue circle highlights the reports within the period ending 2021-09-02 00 UTC (Storm Ida). Panel (d) show the spatial distribution of point (red dots) and gridded (black) impact reports for that same 24-hourly period. Zoomed area shows reports around New York City. 38

4.3 **Percentage (%) of maximum soil saturation for the valid time (VT) corresponding to 2021-09-01 at 00 UTC** Panel (a) shows the percentage of soil saturation in ERA5 reanalysis computed on 2021-08-31 at 18 UTC (t+6). Panels (b) to (f) show ERA5's forecasts computed, respectively, on 2021-08-31 at 00 UTC (t+24) - day 1, 2021-08-30 at 00 UTC (t+48) - day 2, 2021-08-29 at 00 UTC (t+72) - day 3, 2021-08-28 at 00 UTC (t+96) - day 4, and 2021-08-27 at 00 UTC (t+120) - day 5. 45

4.4 **Standard deviation of the filtered sub-grid orography (static field).** The figure presents the map plot showing the values for the standard deviation of the filtered sub-grid orography in ERA5 over the CONUS. 47

4.5 **Leaf area index (climatological field).** Panels (a) to (d) show examples of leaf area index values in ERA5 over the CONUS, respectively, for a day in mid-winter (15th of January), mid-spring (15th of April), mid-summer (15th of July), and mid-autumn (15th of October). 48

4.6 **Graphical representation of the ecPoint post-processing technique, from (Pillosu et al., 2025a).** Panel (a) shows the error formulation for accumulated variables (Forecast Error Ratio, FER) and the error distribution for all cases in the training dataset (Mapping Function, MF). The example pertains to the calibration of 47r3 ECMWF ENS forecasts for 12-hourly rainfall forecasts. Panel (b) shows the univariate approach for ecPoint (U-ecPoint) represented as a "single-leaf" decision tree (DT, within the black circle), while the multivariate approach (M-ecPoint) is represented as a "multiple-leaf" DT (within the grey square). 50

4.7 **Probability (%) of exceeding the 1-year return period in ERA5-ecPoint.** Panel (a) displays probabilities in ERA5-ecPoint reanalysis for the valid time (VT) ending on 2021-09-02 at 00 UTC. Panels (b) to (f) represent the probabilities in ERA5-ecPoint forecasts for the same VT, but for forecasts at day 1 (t+0,t+24), day 2 (t+24,t+48), day 3 (t+48,t+72), day 4 (t+72,t+96), and day 5 (t+96,t+120), respectively. 53

4.8 **Probability (%) of exceeding the 50-year return period in ERA5-ecPoint.** Similar to Figure 4.7, but for probability of exceeding the 50-year return period. . 54

4.9 **Case study analysed over the thesis - Storm Ida.** Panel (a) shows the track followed by Storm Ida. Panels (b) to (e) shows 24-hourly rainfall totals over 2021-08-30, 2021-08-31, 2021-09-01, and 2021-09-02, between 4.30 am and 9.30 am local time. Observations were obtained from https://maps.cocorahs.org/ 56

5.1 **Schematic on how yes- and non-events are defined for rainfall-based flash flood predictions of areas at risk of flash floods.** The point-scale forecast (mm/24h) for each grid box is compared against a specific rainfall threshold. If the forecast exceeds or is equal to the threshold, it is classified as a ”yes-event” (value 1, shown in light green), indicating a risk of flash flooding. Conversely, forecasts below the threshold are classified as ”non-events” (value 0, shown in light brown). The resulting output is a binary field representing the predictions of areas at risk. 62

5.2 **Schematic illustrating the process of defining a verifying rainfall threshold.** Historical rainfall data from observations (in green) or gridded products (in pink) can be used to construct a probability distribution of rainfall intensity. An X^{th} percentile from the tail of this distribution, representing rare and intense events likely to cause flash flooding, is then selected to determine the verifying rainfall threshold. Verifying rainfall thresholds from a distribution constructed from gridded rainfall estimates is typically smaller than those computed using point-scale rainfall estimates. 63

5.3 **Schematic on how the gridded observational field is created from point/polygon impact reports with instantaneous timestamps.** Panel (a) shows the input flash flood impact reports (from NOAA’s Storm Event Database), consisting of point/polygon flash flood reports with instantaneous timestamps. Panel (b) shows the logic followed in the creation of the observational field, involving grouping reports into the considered accumulation period (24-hourly, 00–00 UTC) and spatially mapping them to the considered model grid (reduced Gaussian N320 at 31 km resolution at the equator). Point reports are assigned to the nearest grid-box, whereas polygon reports are assigned to all grid-boxes within the polygon. The total number of reports accumulated in each grid-box is counted. Panel (c) shows that the final observational gridded field is a binary classification, created by assigning a value of 1 to grid-boxes containing at least one report, and 0 otherwise. 66

5.4 **Contingency table for probabilistic forecasts.** Panel (a) shows a schematic on how a contingency table for probabilistic forecasts is built. Panel (b) shows how, when fixing the probability of exceeding the verifying rainfall threshold, it is possible to build a 2x2 contingency table. A series of 2x2 contingency tables is obtained, one per threshold. Panel (c) shows a schematic of the practical construction of the 2x2 contingency table for each grid-box in the considered geographical domain. 67

5.5 **Breakdown verification scores.** Panel (a) shows examples of reliability diagrams for forecasts with perfect, none, and acceptable reliability. An example of reliability diagram for the case of rare events is also shown, including the sharpness diagram. The red square indicates the forecast probabilities with the largest number of cases. Outside the red square, the reliability diagram becomes noisy. Panel (b) is similar, but for ROC curves. 69

5.6 **Overall verification scores for the rainfall-based forecasts of areas at risk of flash flood.** Panel (a) shows the frequency bias (solid lines) for 1-year (in red), 5-year (in purple), 10-year (in light green), 20-year (in cyan), 50-year (in blue), and 100-year return period (in green). The corresponding shaded areas represent the confidence intervals at 99% confidence level. The inset box contains a zoomed-in version of the panel to better show the frequency bias values close to 1 (representing perfect bias). Panel (b) shows the area under the ROC curve. Lead time equal to r" relates to the statistics computed for ERA5-ecPoint *reanalysis*, while lead times from 1 to 5 (in days) relate to ERA5-ecPoint *forecasts*. 73

5.7 **ROC curves for tp >= 1-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint.** Panel (a) shows the ROC curve (red solid line) for the ERA5-ecPoint reanalysis together with the confidence intervals (red shaded area) at 99% confidence level. Panels (b) to (f) refer to ERA5-ecPoint forecasts, for accumulation periods ending in t+24, t+48, t+72, t+96, and t+120, respectively. The dots with the *diamond* symbol refer to the probability threshold at which the frequency bias has the closest value to 1 (i.e., perfectly reliable forecast), while the dots with the *square* symbol show the value of the frequency bias for the lowest probability threshold available in ERA5-ecPoint (i.e., the 99th percentile). 74

5.8 **ROC curves for tp >= 50-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint.** Similar to Figure 5.7. 75

5.9 **Reliability diagrams for tp >= 1-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint.** Panel (a) shows the reliability diagram (red solid line) for the short-range predictions together with the confidence intervals (red shaded area) at 99% confidence level. Panels (b) to (f) refer to the long-range forecasts for accumulation periods ending in t+24, t+48, t+72, t+96, and t+120, respectively. The inset boxes show the corresponding sharpness diagrams. 76

5.10 **Reliability diagrams for tp >= 50-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint.** Similar to Figure 5.9. 77

6.1 **Overview of the three ensemble data-driven model architectures used in this study.** *Bagging*, with random forests (with XGBoost and LightGBM implementations in random forest mode), *boosting*, with gradient boosting (with XGBoost, LightGBM, and CatBoost implementations in gradient boosting mode), and *neural networks*, with feed-forward architectures (implemented using Keras with TensorFlow backend). 86

6.2 **Workflow for the repeated nested cross-validation.** The outer cross-validation loop utilises Scikit-Learn's "RepeatedStratifiedKFold" function to create k_outer = 5 outer folds (grey blocks) across n_repeats = 1 iterations. Each *outer fold* maintains the class distribution of the *training dataset*, and it is split into an outer training dataset (80%, blocks in shades of pink and orange) and an *outer test dataset* (20%). Within each outer fold, a Bayesian hyperparameter tuning is performed employing the Optuna library through an inner cross-validation procedure over n_trial = 20 repetitions. Each trial evaluates candidate hyperparameters by training on *inner training folds* and validating on *inner validation folds*, with performance measured as the mean AUC-ROC or AUC-PR. The optimal hyperparameter set, identified by maximising the selected evaluation metric, is used to train the final model on the complete outer training subset. Model performance is assessed on the held-out *outer test fold* using AUC-ROC and AUC-PR. The best-performing fold is retrained on the original *training dataset* for operational deployment. Independent, more extensive verification of the data-driven predictions is performed using the *verification dataset*, considering the Precision-Recall curve and AUC-PR, the ROC curve and AUC-ROC, reliability diagrams, and frequency bias. 91

6.3 **Schematic on how yes- and non-events are defined for the forecasts of probability of flash flood (in %, at grid-scale).** The forecast of probability of flash flood (%) for each grid box is compared against a corresponding probability threshold (optimised on the F1-score). If the forecast exceeds or is equal to the threshold, it is classified as a "yes-event" (value 1, shown in light green), indicating a risk of flash flooding. Conversely, forecasts below the threshold are classified as "non-events" (value 0, shown in light brown). The resulting output is a binary field representing the predictions of areas at risk. 94

6.4 **Breakdown score to assess discrimination ability for imbalanced training datasets: Precision-Recall curves.** Examples of Precision-Recall curves for forecasts with perfect, none, and good discrimination ability (for balanced datasets). The figure also shows the typical precision-recall curves for imbalanced datasets, with ideal (i-a) and good (i-b) discrimination ability. 95

6.5 **Optuna's hyperparameter optimisation history.** Evolution of the two evaluation metrics (AUC-ROC, panels (a) to (l) - and AUC-PR, panels (m) to (x)) maximised during the 20 trials run over the *inner validation folds* to tune the hyperparameters of six data-driven models (from top to bottom): random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. The lines in shades of grey indicate individual trial performances, whilst lines in shades of orange highlight the best-performing hyperparameter set, identified by Optuna's Bayesian optimisation process. The shades of grey and orange represent the values of the evaluation metrics for each outer fold (lightest shade for the first outer fold and darkest for the latest). Panels (a) to (f) and (m) to (r) represents the results obtained using the standard binary cross-entropy loss functions - mostly used for balanced datasets - whilst panels (g) to (l) and (s) to (x) present the outcomes obtained with the weighted loss functions (specifically configured for imbalanced data). 97

6.6 **Optuna's training time.** Evolution of training times (in seconds) for each k_outer=5 outer folds across the n_trials = 20 trials run over the *inner training folds* (the solid line represents the mean while the shaded area represents the minimum and maximum values). The training times for six data-driven models (from top to bottom) are shown: random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. Training times are shown for both evaluation metrics (AUC-ROC - panels (a) to (l) - and AUC-PR - panels (m) to (x)) and loss function configurations (balanced and imbalanced datasets). Inset plots provide magnified views where appropriate. 99

6.7 **Model generalisation from nested cross-validation** Values of the two evaluation metrics (AUC-ROC, panels (a) to (l) - and AUC-PR, panels (m) to (x)) across the 20 trials run over the *inner validation folds* (the solid line represents the mean while the shaded area represents the minimum and maximum values) and the *outer test fold* (the solid line correspond to the single realisation per outer fold). Panels (a) to (f) and (m) to (r) represent the results obtained using the standard binary cross-entropy loss functions - mostly used for balanced datasets - whilst panels (g) to (l) and (s) to (x) present the outcomes obtained with the weighted loss functions (specifically configured for imbalanced data). 101

6.8 **Optuna's hyperparameter importance for the XGBoost implementation of gradient boosting.** Normalised absolute Pearson's correlation coefficients obtained for the n_trials = 20 trials run over the *inner training folds* (bars represent mean values, whilst error bars show the minimum and maximum values across the Optuna trials). Feature importance is shown for models trained with loss functions for balanced datasets and specific for imbalanced datasets, and for both evaluation metrics (AUC-ROC and AUC-PR). 102

6.9 **Optuna's hyperparameter importance for the LightGBM implementation of gradient boosting.** Similar to Figure 6.8 103

6.10 **Optuna's hyperparameter importance for the CatBoost implementation of gradient boosting.** Similar to Figure 6.8 104

6.11 **Optuna's hyperparameter importance for the XGBoost implementation of random forest.** Similar to Figure 6.8 105

6.12 **Optuna's hyperparameter importance for the LightGBM implementation of random forest.** Similar to Figure 6.8 106

6.13 **Optuna's hyperparameter importance for feed-forward neural network.** Similar to Figure 6.8 107

6.14 **Verification results: overall scores** The first, second, and third columns show, respectively, the estimates for the area under the ROC curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the frequency bias (FB) for the six considered data-driven models. The estimates of the overall verification scores are shown for the training dataset and the verification dataset. Results are presented for the models trained considering both types of loss functions and evaluation metrics. 108

6.15 **Verification results: breakdown scores (ROC curves)** ROC curves computed for the training dataset and the verification dataset, for six data-driven models (from top to bottom): random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. The solid lines represent ROC curves computed considering forecast probabilities discretised every 1%, whilst the dashed lines represent ROC curves computed using a finer discretisation of probabilities (0.01%). ROC curves are shown for the two evaluation metrics (AUC-ROC - panels (a) to (l) - and AUC-PR - panels (m) to (x)) and types of loss function configurations (balanced and imbalanced datasets) considered during the training of the data-driven models. 109

6.16 **Verification results: breakdown scores (Precision-Recall curves)** Similar to Figure 6.15 but for the Precision-Recall curve 111

6.17 **Verification results: breakdown scores (Reliability diagrams)** Similar to Figure 6.15 but for reliability diagrams. 112

6.18 **Verification results for medium-range forecasts for the XGBoost implementation of gradient boosting (trained with the loss function for balanced data and hyperparameters maximised for AUC-ROC).** Panels (a) to (c) show the overall scores, respectively, for the area under the ROC curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the frequency bias (FB). The remaining panels show the breakdown scores, respectively, for the ROC curves (Panels (d) to (f)), the Precision-Recall curves (Panels (g) to (i)), and for the Reliability Diagram (Panels (j) to (l)). 114

6.19 **SHAP ((SHapley Additive exPlanations)) values over the verification dataset for the XGBoost implementation of gradient boosting (trained with the loss function for balanced data and hyperparameters maximised for AUC-ROC).** Panel (a) shows the global feature importance ranking (most important features in descending order). Panels (b) to (d) show the dependency plots between *tp_prob_1* and, respectively, *LAI*, *sdfor*, and *swvl*. Panels (e) to (g) show the same, but for *tp_prob_50.* 115

6.20 **Areas at risk of flash floods.** Probability of having a flash flood in a grid-box, valid for the 24-hourly accumulation ending on 2021-09-02 at 00 UTC. Panel (a) shows the probabilities computed with reanalysis data, while panels (b) to (f) show the probabilities computed with forecast data, respectively for day 1 (t+0,t+24), day 2 (t+24,48), day 3 (t+48,t+72), day 4 (t+72,t+96), and (t+96,t+120).117

7.1 **Training approaches adopted in the sensitivity analysis.** Panel (a) describes Training Approach 1 (TA1) - where the flash flood reports are randomly reduced uniformly over the whole domain by 10% (TA1-1), 50% (TA1-2), and 90% (TA1-3), and during training, the model sees the full domain. Panel (b) describes Training Approach 2 (TA2) - where the flash flood reports are present only over one part of the domain (TA2-1 for reports in the west and TA2-2 in the East), but during training, the model still sees the full domain. Panel (c) describes Training Approach 3 (TA3) - where the flash flood reports are present only over one part of the domain (TA3-1 for reports in the west and TA3-2 in the East), and during training, the model sees only the part of the domain with reports. 127

7.2 **Objective verification: overall scores (AUC-ROC, AUC-PR, FB)** Panel (a) shows the area under the ROC curve (AUC-ROC), computed for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets and hyperparameters optimised to maximise AUC-ROC (as developed in Chapter 6). The score is computed with data from the verification dataset (from 2021 to 2024). The solid line shows the scores computed using the three considered training approaches (TA1-3) as described in Figure 7.1. The dashed line represents the score obtained when the model was trained with the full training dataset (refer to Figure 6.14a. Panels (b) and (c) show, respectively, the area under the precision-recall curve and the frequency bias. Refer to Figures 6.14e and 6.14i for the corresponding scores obtained when the model was trained with the full training dataset. 130

7.3 **Objective verification: breakdown scores (ROC curves)** ROC curves are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.15c). Panel (a) shows the ROC curve for the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the ROC curves for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the ROC curves obtained for TA2, while panels (g) and (h) show the ROC curves obtained for TA3. The ROC curves drawn with solid lines correspond to a probability discretisation of 1%, with probabilities of exceedance ranging from 0 to 99%. The ROC curves drawn with dashed lines correspond to a probability discretisation of 0.01%, with probabilities of exceedance ranging up to values that depend on the TA, and shown in Figure 7.1. 132

7.4 **Objective verification: breakdown scores (Precision-Recall curves)** Precision-Recall curves are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.16c). Panel (a) shows the precision-recall curve for the model trained with the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the precision-recall curves for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the ROC curves obtained for TA2, while panels (g) and (h) show the ROC curves obtained for TA3. 134

7.5 **Objective verification: breakdown scores (Reliability Diagrams)** Reliability diagrams are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.17c). Panel (a) shows the reliability diagram for the model trained with the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the reliability diagrams for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the reliability diagrams obtained for TA2, while panels (g) and (h) show the reliability diagrams obtained for TA3. 135

7.6 **Map plots of probabilities for areas at risk of flash flood for different training approaches, for reanalysis data.** Panel (a) shows the baseline probabilities obtained by training the considered XGBoost model (with balanced loss function and hyperparameters optimised by maximising the AUC-ROC metric) with the full training dataset as shown also in Figure 6.20a, in Chapter 6. Panel (b) to (d) show the probabilities of areas at risk of flash floods for TA1, respectively, TA1-1, TA1-2, and TA1-3. Panels (e) and (f) show the probabilities for TA2, respectively, TA2-1, and TA2-2. Panels (g) and (h) show the probabilities for TA3, respectively for TA3-1 and TA3-2. 137

7.7 **Flash floods in Spain in October 2024.** The valid time (VT) for all plots is from the 29th of October 2024 at 00 UTC to the 30th of October 2024 at 00 UTC. Panel (a) shows rainfall observations (mm/24h) over Valencia, taken from Gascón et al. (2025). Panels (b) and (c) show the ERA5-ecPoint rainfall probabilities (%) of exceeding, respectively, the 1- and 50-year return period from reanalysis over Europe. Panel (d) show the probability (%) of flash flood computed with reanalysis data for a zoomed-in area over Spain. Panels (e) to (i) show the probability of flash flood computed from ERA5-ecPoint rainfall forecasts (FC), respectively, for day 1 - 2021/07/20 at 00 UTC (t+0, t+24) - day 2 - 2021/07/19 at 00 UTC (t+24, t+48) - day 3 - 2021/07/18 at 00 UTC (t+48, t+72) - day 4 - 2021/07/17 at 00 UTC (t+72, t+96) - and day 5 - 2021/07/16 at 00 UTC (t+96, t+120). 139

7.8 **Flash floods in China in July 2021.** The valid time (VT) for all plots is from the 20th of July 2021 at 00 UTC to the 21st of July 2021 at 00 UTC. Panel (a) shows rainfall observations (mm/24h). Panels (b) and (c) show the ERA5-ecPoint rainfall probabilities (%) of exceeding, respectively, the 1- and 50-year return period from reanalysis. Panel (d) show the probability (%) of flash flood computed with reanalysis data. The red circle highlights the high-to-extreme rainfall totals in the area surrounding and over Zhengzhou, the blue circle highlights the high rainfall totals over Hong Kong and surroundings, while the purple circle highlights an area with moderate rainfall over Central Mongolia. Panels (e) to (i) show the probability of flash flood computed from ERA5-ecPoint rainfall forecasts (FC), respectively, for day 1 - 2021/07/20 at 00 UTC (t+0, t+24) - day 2 - 2021/07/19 at 00 UTC (t+24, t+48) - day 3 - 2021/07/18 at 00 UTC (t+48, t+72) - day 4 - 2021/07/17 at 00 UTC (t+72, t+96) - and day 5 - 2021/07/16 at 00 UTC (t+96, t+120). These plots focus on the area surrounding and over Zhengzhou (within the red circle). 141

LIST OF TABLES

4.1 **Features used from the Storm Event Database.** The table presents the name of the features, called ”Keys” in the database, the variable types (e.g., string, data object, or float), their units, and description. 40

4.2 **ERA5 reanalysis and forecasts.** The table describes the characteristics of the ERA5 reanalysis and forecasts. 42

4.3 **Parameters used from ERA5 and ERA5-ecPoint.** Description of the parameters used to compute the hydrological and static features used in the developed data-driven models to identify and predict areas at risk of flash floods. 43

4.4 **Maximum soil saturation** The table shows the pre-defined values of maximum soil saturation per soil as suggested by Balsamo et al. (2009). 44

5.1 **Severity of flash-flood-triggering rainfall events**. The table shows the percentiles that define different severity categories for flash-flood-triggering rainfall events and their equivalent in return periods (in years, for mm/24h). Results for the starred (*) return periods are discussed but not shown in the ”Results” Section 5.3. 65

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AUC-PR	Area Under the Precision-Recall Curve
AUC-ROC	Area Under the Relative Operating Characteristic Curve
CAMELS	Catchment Attributes and MEteorology for Large-sample Studies
CN	Correct Negative
CNN	Convolutional Neural Network
CONUS	CONtiguous United States (of America)
DEM	Digital Elevation Model
DesInventar	Disaster Inventory System
ECMWF	European Centre for Medium-range Weather Forecasts
ecPoint	Statistical post-processing system that creates probabilistic forecasts at point-scale
EFAS	European Flood Alert System
EM-DAT	Emergency Events Database
ENS	ECMWF ENSemble
ERA5	European Reanalysis Version 5
ERICHA	European Rainfall-InduCed Hazard Assessment
ESSL	European Severe Storms Laboratory

ESWD	European Severe Weather Database
FA	False Alarm
FAR	False Alarm Rate
FB	Frequency Bias
FER	Forecast Error Ratio
FFG	Flash Flood Guidance
FLASH	
GloFAS	Global Flood Alert System
GPCP	Global Precipitation Climatology Project
H	Hit
HR	Hit Rate
IFS	Integrated Forecasting System
IoT	Internet of Things
IPCC	Intergovernmental Panel on Climate Change
KGE	Kling–Gupta efficiency
LAI	Leaf Area Index
LSTM	Long-Short Term Memory
M	Miss
MF	Mapping Function
ML	Machine Learning
MSWEP	Multi-Source Weighted-Ensemble Precipitation
NOAA	National Oceanic and Atmospheric Administration
NSE	Nash-Sutcliffe Efficiency coefficient
NWP	Numerical Weather Prediction
NWS	National Weather Service

PT	Probability Threshold
RC	Research Challenge
RG	Research Gap
RNN	Recurrent Neural Network
RO	Research Objective
ROC	Receiver Operating Characteristic
RQ	Research Question
SED	Storm Event Database
UN	United Nations
VT	Valid Time
WMO	World Meteorological Organization

CHAPTER 1

GENERAL INTRODUCTION

The socio-economic impacts of flash floods

Flash floods represent the deadliest and most devastating hazards, causing over 5,000 fatalities annually and accounting for ∼85% of global flood incidents (Dordevic et al., 2020). The impact of flash floods extends across urban and rural landscapes, profoundly affecting lives, livelihoods, and critical infrastructure worldwide (Liu et al., 2024). The socio-economic (Ebi et al., 2021) and environmental (Zhang et al., 2024) impacts of flash

DEFINITIONS CONSIDERED IN THIS THESIS

Flash Flood

Two types of flash floods are considered: *fluvial flash floods*, occurring in catchments between 100 and 500 km^2, are defined as a rapid water level rise in a stream or creek above a predetermined flood level (NWS, 2025); and *pluvial flash floods*, occurring in catchments less than 100 km^2, are defined as flooding resulting from rainfall when water ponds or flows over the ground before it enters a natural or man-made drainage system or watercourse, or when it cannot enter because the system is already full to capacity (Speight et al., 2021). For both types of flash floods, flooding is considered to occur within a few hours of the causative event. In this thesis, the only considered causative event is rainfall.

Area at Risk of Flash Floods

It refers to a polygon - or area - that might experience flash flooding due to the expected rainfall.

floods can be severe and transcend the traditional divide between developed and developing countries. In October 2024, flash floods in Valencia, Spain, claimed more than 200 lives and caused extensive damage in 87 municipalities (Grau-Bove et al., 2024), while between March and September 2024, sustained periods of intense rainfall and subsequent flash flooding in Pakistan and Afghanistan resulted in 1084 deaths, 2600 injuries, and extensive damage to houses (Wikipedia, 2025). In low- and middle-income countries in Africa, Latin America, and Asia, flash floods can also exacerbate existing socio-economic and environmental vulnerabilities to the extent of displacing entire populations (Stephens and Levi, 2024) or undermining food security and food safety (Agabiirwe et al., 2022; Duchenne-Moutien and Neetoo, 2021). Impacts from flash floods are more severe primarily due to rapid, unregulated urbanisation in flood-prone areas, limited infrastructure for flood management, insufficient early warning systems, and economic constraints on implementing preventive measures (Douglas, 2017; Pinos and Quesada-Román, 2022; Wang et al., 2021). Regions affected by flash floods can also be vulnerable to waterborne diseases, with outbreaks of cholera, typhoid, and other infectious diseases occurring when floodwaters contaminate drinking water sources or overwhelm sanitation systems (Lee et al., 2020). Populations affected by extremely severe flash floods may also experience serious psychological impacts, including anxiety, depression, and post-traumatic stress disorder (Iqbal et al., 2023).

On the urgent need for accurate, timely, and scalable flash flood forecasts

As climate change increases the frequency and intensity of extreme rainfall (WMO, 2025; IPCC, 2023), also in historically low-risk regions (Fowler et al., 2021), a comprehensive re-evaluation of risk management, adaptation, and mitigation frameworks is needed to protect vulnerable communities. Recognising their severe and growing impacts, WMO targets flash floods as one of its top priority natural hazards (WMO, 2025). The UN's 'Early Warnings for All' initiative, launched in 2022, which aims to protect every person on Earth with early warning systems by 2027, places flash floods at the forefront of its agenda (UN, 2022). Forecasts with global coverage that are accurate and timely (e.g. several days in advance) are crucial to the success of such an initiative as they could enable targeted protective decisions worldwide (Merz et al., 2020), including in regions where longer lead times are crucial for mobilising resources and executing emergency plans (Bazo et al., 2019).

Current challenges in flash flood forecasting: scaling predictions for large domains and extending lead times

Despite general advances in flash flood prediction (for example, the development of high-resolution physical and data-driven NWP and hydro-

Figure 1.1: Global Implementation of WMO's Flash Flood Guidance System (FFGS). The map shows in blue the countries that have adopted the FFGS as indicated in (Georgakakos et al., 2022) and the WMO's Global FFGS Status at https://experience.arcgis.com/experience/bb4357054fb5475a9a4cf688c4180454/page/FFGS-Status?views=Flash-Flood-Component%2CFlash-Flood-Component .

logical models), significant technical and methodological obstacles persist for the development of medium-range flash flood forecasts (i.e. up to 5 days ahead), over a continuous global domain (Zanchetta and Coulibaly, 2020). Such obstacles include inherent uncertainty in extreme, localised rainfall forecasts beyond a few hours, computational demands of running high-resolution models over large domains, and the limited availability of real-time hydro-meteorological observations. These limitations hinder the ability to predict flash floods with sufficient lead time and across a large-scale, continuous domain (e.g., global), leaving many regions unprotected, i.e. without early warning systems (Al-Rawas et al., 2024). The WMO's Flash Flood Guidance System attempts to address the patchy spatial coverage of flash flood forecasting systems (Georgakakos et al., 2022). This initiative, however, focuses on implementing distinct regional systems around the world, not resolving the issue of spatial patchy coverage (Figure 1.1). Furthermore, the system's reliance on high-density observational networks, km-scale NWP model outputs, and high computational costs compromises the scalability of the project in regions with limited resources. Therefore, developing robust medium-range flash flood forecasting capabilities on a global scale remains one of the pressing challenges in modern hydrology.

Opportunities for global medium-range flash flood forecasts: proven effectiveness of index-based systems over large-scale domains, enhanced quality of medium-range global NWP rainfall predictions, and emergence of data-driven approaches for hydrological applications

The unprecedented convergence of three key scientific advancements over the last decade has created a unique opportunity to develop medium-range flash flood forecasts with global coverage. First, index-based flash flood forecasting systems, focusing on key variables such as rainfall and soil moisture, have proven more effective and computationally efficient at national and continental scales than complex physically-based models (Alfieri and Thielen, 2015). However, their reliance on high-resolution, short-range rainfall forecasts from radars or km-scale NWP models still limits their spatial coverage to data-rich regions like Europe and the US, and restricts forecasts lead times to nowcasting time scales (i.e. a few hours), thereby reducing available preparedness and response time (Luong et al., 2021; Maybee et al., 2024). Second, over the past decade, global (ensemble) NWP models have significantly improved their ability to forecast extreme rainfall up to the medium-range lead times (Lavers et al., 2021; Haiden et al., 2023). Despite their coarse spatial resolution (typically >10 km), there is growing interest in testing global NWP forecasts for flash flood applications and extending prediction lead times (Bucherie et al., 2022b). Moreover, statistical post-processing techniques make these predictions more palatable for flash flood forecasting (Vannitsem et al., 2021). Third, the recent success of data-driven approaches in predicting riverine floods (Nearing et al., 2024) has increased the interest in extending their application to flash flood forecasting. Notwithstanding the innovative paradigm of *training a model where data is available and applying it globally* (Kratzert et al., 2024), the paucity of observational data suitable for flash flood modelling continues to hinder the development of data-driven approaches for large-scale flash flood forecasting (Alzubaidi et al., 2023). When run using global medium-range NWP model outputs, simpler machine learning models (e.g., decision-tree-based algorithms or feed-forward neural networks), optimised for sparse and imbalanced datasets[1], and informed by physical insights from index-based models, may enable the development of a proof-of-concept for medium-range data-driven hydro-meteorological predictions of areas at risk of flash floods with true global coverage.

[1]Sparse, imbalanced data describe datasets where most entries are zero or missing (sparse) and the target classes occur at very unequal frequencies (imbalanced), often biasing models toward the majority class.

1.1 Research questions and objectives, and contributions to knowledge

Development of flash-flood-focused verification framework and establishment of rainfall-based performance benchmark for comparative assessment against more sophisticated predictions of areas at risk of flash floods

While new global NWP rainfall forecasts are regularly developed, their effectiveness in identifying areas at risk of flash floods remains largely untested. Most verification efforts compare predicted rainfall against rainfall observations, implicitly assuming that improved rainfall forecasts will translate into better flash flood prediction capabilities (Gascón et al., 2024). However, this assumption requires direct validation through a flash-flood-focused assessment.

> Research Question n.1 (RQ1): Can post-processed global NWP rainfall forecasts successfully identify areas at risk of flash floods up to medium-range lead times?

To address RQ1, the first research objective of this thesis involves departing from the traditional rainfall-to-rainfall verification approach and adopting, instead, a *flash-flood-focused verification framework* that directly compares rainfall forecasts with flash flood impact reports - to encompass fluvial and pluvial flash flood events. This framework must overcome two main methodological challenges: addressing spatio-temporal uncertainties in flash flood reports, and establishing meaningful performance measures that account for the inherent rarity of flash flood events and the differences between continuous rainfall predictions and binary flash flood occurrence (as measured by impact reports). Addressing RQ1 delivers two fundamental contributions to knowledge. First, it establishes a performance baseline by evaluating state-of-the-art global NWP rainfall forecasts against flash flood impact reports, providing a quantitative benchmark against which more sophisticated predictions (e.g., incorporating hydro-meteorological parameters) can be assessed. Second, the framework itself constitutes a contribution - a standardised assessment tool applicable to any flash flood prediction, whether rainfall-based or hydro-meteorological, physical or data-driven. This versatile framework will underpin all subsequent verification analyses in this thesis, enabling systematic comparison across increasingly complex prediction methodologies.

Development of data-driven hydro-meteorological predictions of areas at risk of flash floods up to medium-range lead times

Data-driven models (Nearing et al., 2024) and large-sample hydrology (Kratzert et al., 2024) have recently transformed riverine flood prediction, yet data-driven flash flood prediction remains confined at catchment/regional or (Song et al., 2020; Saleh et al., 2024; Zhao et al., 2025) or national

scale (Zhao et al., 2022), and forecast lead times rarely exceed 24 hours. The development of medium-range, global data-driven flash flood prediction systems faces a fundamental obstacle: severe class imbalance between flash flood and non-flood events in observational datasets for this hazard. This severe imbalance - stemming from the inherent rarity of flash flood events, the predominant number of ungauged flashy catchments, and the systematic underreporting of this hazard in global impact databases (Panwar and Sen, 2020; Kratzert et al., 2023; Jonkman et al., 2024).

> Research Question n.2 (RQ2): Are medium-range data-driven hydro-meteorological predictions of areas at risk of flash floods feasible[2], with skill quantified relative to the rainfall-only baseline established in RQ1. with global reanalysis, forecasts, and impact flash flood reports?

To address RQ2, the second research objective of this thesis involves developing data-driven models that integrate hydro-meteorological variables from global reanalysis and global medium-range NWP forecasts, and flash flood impact reports to predict areas at risk of flash floods, from short- (i.e., day 1) to medium-range lead times (i.e., day 5). This development must overcome three methodological challenges: selecting architectures that handle severe class imbalance effectively - where parsimonious approaches, such as regularised ensemble methods and shallow neural networks, may outperform deep learning (Kumar et al., 2021; Xu et al., 2022; Luo et al., 2025); feature engineering that balance informativeness, interpretability, and computational efficiency when identifying key hydro-meteorological variables; building ensemble models to quantify uncertainty and provide reliable risk estimates (Saleh et al., 2024). Addressing RQ2 delivers two fundamental contributions to knowledge. First, it establishes the feasibility and predictability limits of medium-range data-driven predictions of areas at risk of flash floods. Second, it demonstrates whether more sophisticated predictions, integrating hydro-meteorological parameters and complex probabilistic data-driven approaches, enhance the predictive ca-

[2] In this thesis, "feasible" is used in a combined sense, encompassing both (i) technical feasibility (i.e., the practical possibility of developing and running the modelling framework with the available global reanalysis, medium-range NWP predictions, and flash flood impact reports), and (ii) good predictive performance (in terms of reliability and discrimination ability) up to medium-range lead times (up to day 5) of forecasts that do not collapse to the majority (i.e., non-flash-flood-event).

pability of identifying areas at risk of flash floods up to medium-range lead times compared to the rainfall-only baseline established in RQ1.

Systematic empirical evaluation of training strategies for global predictions of areas at risk of flash floods under varying spatial coverage and data density scenarios.

The verification analysis and data-driven model development presented in the previous paragraphs focus on a regional domain with high-quality, high-density flash flood impact reports. Even though these regionally-trained data-driven models may demonstrate strong performance within the considered domain, this thesis aims to develop a proof-of-concept for predicting areas at risk of flash floods *over a continuous global domain*. The lack of high-density global flash flood impact databases (Panwar and Sen, 2020) creates a fundamental tension between two contrasting approaches for developing predictions of areas at risk of flash floods with a continuous global coverage: training models on sparse but global datasets versus leveraging high-quality regional observations and applying the trained model globally, as suggested by Kratzert et al. (2024).

> Research Question n.3 (RQ3): How does coverage-density trade-off influence training data strategies to develop predictions of areas at risk of flash floods over a continuous global domain?

To address RQ3, the third research objective of this thesis involves the assessment - through a systematic empirical sensitivity analysis - of how varying spatial coverage and data density scenarios may influence training strategies when creating global predictions with regionally-trained data-driven models (as those developed to address RQ2). This investigation must overcome two primary challenges: quantify how performance degrades across different sensitivity analysis configurations to identify the optimal trade-off between geographical coverage and data quality, and establish subjective validation protocols for regions lacking comprehensive flash flood databases. Addressing RQ3 delivers two fundamental contributions to knowledge. First, it provides empirical evidence quantifying how different training data strategies — from sparse global coverage to dense regional coverage — may influence the accuracy of global predictions. Second, it assesses whether regionally-trained models can maintain meaningful predictive skill when applied to data-scarce regions outside the training domain. This evidence supports a pragmatic pathway toward global flash flood early warning systems, particularly valuable for regions where traditional forecasting approaches face data or resource limitations. Moreover, the outcomes of this research would align with the UN's "Early

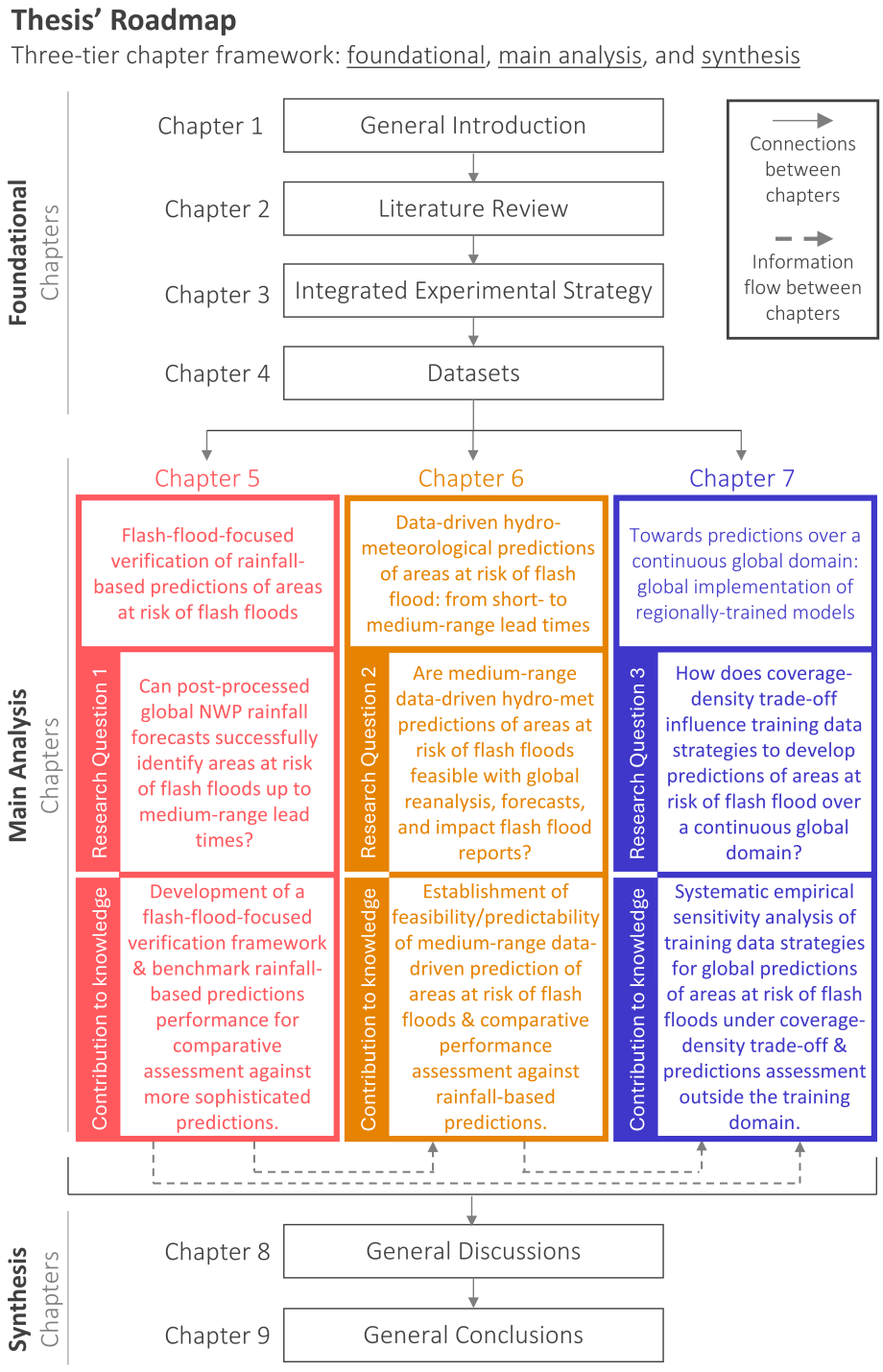

Figure 1.2: Thesis' roadmap. Three-tier chapter framework comprising **Foundational** Chapters (Chapter 1-4: General Introduction, Literature Review, Integrated Experimental Strategy, and Datasets), **Main Analysis** Chapters (Chapter 5: Development of a flash-flood-focused verification framework; Chapter 6: Development of medium-range data-driven hydro-meteorological predictions of areas at risk of flash floods; Chapter 7, blue: Global implementation of regionally-trained data-driven models), and **Synthesis** Chapters (Chapter 8-9: General Discussions and General Conclusions). Research questions and main output(s) are indicated for each main analysis chapter.

Warnings for All" initiative by potentially extending life-saving warnings to historically underserved communities.

1.2 Thesis structure

The thesis is organised in nine chapters.

Chapter 1 established the foundation for this research by addressing the critical need for improved flash flood early warning systems worldwide. The chapter presented the three interconnected research questions and objectives addressed in this thesis.

Foundational Chapter - General Introduction

Chapter 2 presents a synthesis of the scientific literature in flash flood prediction, establishing the theoretical foundations that delineate the methodological gaps addressed by the three research questions, which were presented in Section 1.1.

Foundational Chapter - Literature Review

Chapter 3 delineates the integrated experimental strategy through which the three research questions are systematically addressed, demonstrating their dependencies and relationships.

Foundational Chapter - Integrated Experimental Strategy

Chapter 4 presents the observational and modelled (reanalysis and forecasts) datasets used to address the three research questions presented in Section 1.1. In this chapter, the strengths and weaknesses of each dataset are presented, as well as a discussion on how those might impact the results obtained in the main analysis chapters (5 to 7).

Foundational Chapter - Datasets

Chapter 5 addresses RQ1 by developing a flash-flood-focused verification framework - a standardised assessment tool applicable to diverse predictions of areas at risk of flash floods - and establishes a quantitative performance baseline for rainfall-based predictions against which more sophisticated predictions may be compared.

Main Analysis Chapter - Flash-flood-focused verification of rainfall-based predictions of areas at risk of flash floods

Chapter 6 addresses RQ2 by developing and evaluating multiple data-driven architectures trained on hydro-meteorological reanalysis data and run with medium-range global NWP hydro-meteorological forecasts to analyse forecast predictability. The chapter also demonstrates whether integrating additional variables enhances forecast performance compared to the rainfall-based benchmark.

Main Analysis Chapter - Data-driven hydro-meteorological predictions of areas at risk of flash flood: from short- to medium-range lead times

Main Analysis Chapter - Towards predictions over a continuous global domain: global implementation of the regionally-trained models

Chapter 7 addresses RQ3 by providing empirical evidence for optimal training data strategies to expand regional training to create predictions of areas at risk of flash floods over a continuous global domain, as well as assessing coverage-density trade-offs through systematic sensitivity analysis.

Synthesis Chapter - General Discussions

Chapter 8 synthesises and discusses the research findings in the main analysis chapters, critically evaluating how effectively each research question was addressed. The chapter examines the broader implications of the thesis's outcomes for flash flood risk reduction and emergency management, exploring how these advancements can enhance global early warning capabilities against the impacts of flash floods.

Synthesis Chapter - General Conclusions

Chapter 9 concludes the thesis by stressing its novel contributions to flash flood prediction, in both scientific advances and forecasting practice. It provides final assessments for each research question, acknowledging methodological limitations whilst clearly stating the scientific advancements achieved. The chapter concludes by exploring future research directions that could strengthen the use of forecasts from global NWP models and data-driven approaches for predicting flash floods over a continuous global domain and further extending forecast lead times.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the scientific foundations underpinning the development of medium-range predictions of areas at risk of flash floods with global coverage. The chapter is organised around three interconnected themes. Section 2.1 explores the extent to which medium-range global NWP rainfall forecasts can successfully identify areas at risk of flash floods. Section 2.2 reviews the role of data-driven approaches in enabling medium-range predictions of areas at risk of flash floods. Section 2.3 explores training data strategies for overcoming data-scarcity constraints when scaling regionally-trained data-driven models to continuous global coverage. Each section moves from established achievements through current challenges to opportunities, thereby clarifying the rationale for the analyses presented in the main chapters of the thesis.

2.1 Can global medium-range NWP rainfall forecasts successfully identify areas at risk of flash floods?

Flash floods and their challenging prediction

Flash floods are characterised by a rapid hydrological response to intense rainfall events, with runoff timescales ranging from mere minutes to a few hours (Davis, 2001). While the interaction with local topography, land use, and drainage systems modulates flash flood severity (Marchi et al., 2010; Villaça et al., 2025), localised extreme rainfall is the primary driver of

flash floods[1] (Schumacher, 2017; Borga et al., 2014; Archer and Fowler, 2018; Meyer et al., 2022). Unlike riverine floods that develop gradually over many hours or days from widespread rainfall across large watersheds (Wohl and Lininger, 2022), flash floods generate in small catchments (typically under 100 km^2 or 500 km^2) and within minutes to hours after the triggering rainfall event (Braud et al., 2014; Blöschl et al., 2015). Hence, accurately characterising and predicting localised extreme rainfall (i.e., its intensity and spatial distribution) determines our ability to identify areas at heightened risk of flash flooding. Near-perfect precision in forecasting the correct distribution of rainfall totals at the precise locations is required, as any underestimation or spatial misplacement, even by a few kilometres, can leave the interested catchment virtually dry and eliminate any flash-flood signal (Nicótina et al., 2008; Douinot et al., 2016; Clark et al., 2016; Wang and Karimi, 2022).

The most important meteorological ingredients for flash flood-producing storms: moisture supply, uplift of moist air, and slow-moving convective systems

The key meteorological ingredients that contribute in the generation of flash flood-producing storms are (1) the ample and persistent moist supply, (2) the uplift of the moist air, and (3) the presence of mechanisms that cause precipitation to occur continuously or repeatedly over the same area (Doswell et al., 1996). Sufficient moisture availability typically depends on sustained transport from oceanic or continental evaporative sources via low-level jets, atmospheric rivers, or persistent onshore flow (Figure 2.1a). The magnitude of precipitable water in the atmospheric column sets an upper bound on rainfall intensity for a given storm duration. The uplift of this moist air to altitudes where adiabatic cooling induces condensation is achieved through four main mechanisms (Figure 2.1b). Convective uplift occurs when the sun heats the surface, generating a thermal instability that uplifts moist air to the condensation level. When horizontally moving air encounters mountains, it is forced to ascend, cooling adiabatically and causing the moisture contained within the air mass to condense. Consequently, precipitation typically occurs on the windward slope, whilst rainfall on the leeward side remains limited due to the descent of moisture-depleted air (rain shadow). Frontal uplift arises where warmer, less dense air ascends over a cooler, denser air mass. Low-level convergence associated with cyclonic systems

[1]While acknowledging that flash floods can also be triggered by other mechanisms such as ice jams, dam breaks, and landslides, this thesis focuses exclusively on rainfall-triggered flash floods, as they represent the most frequent cause of these events.

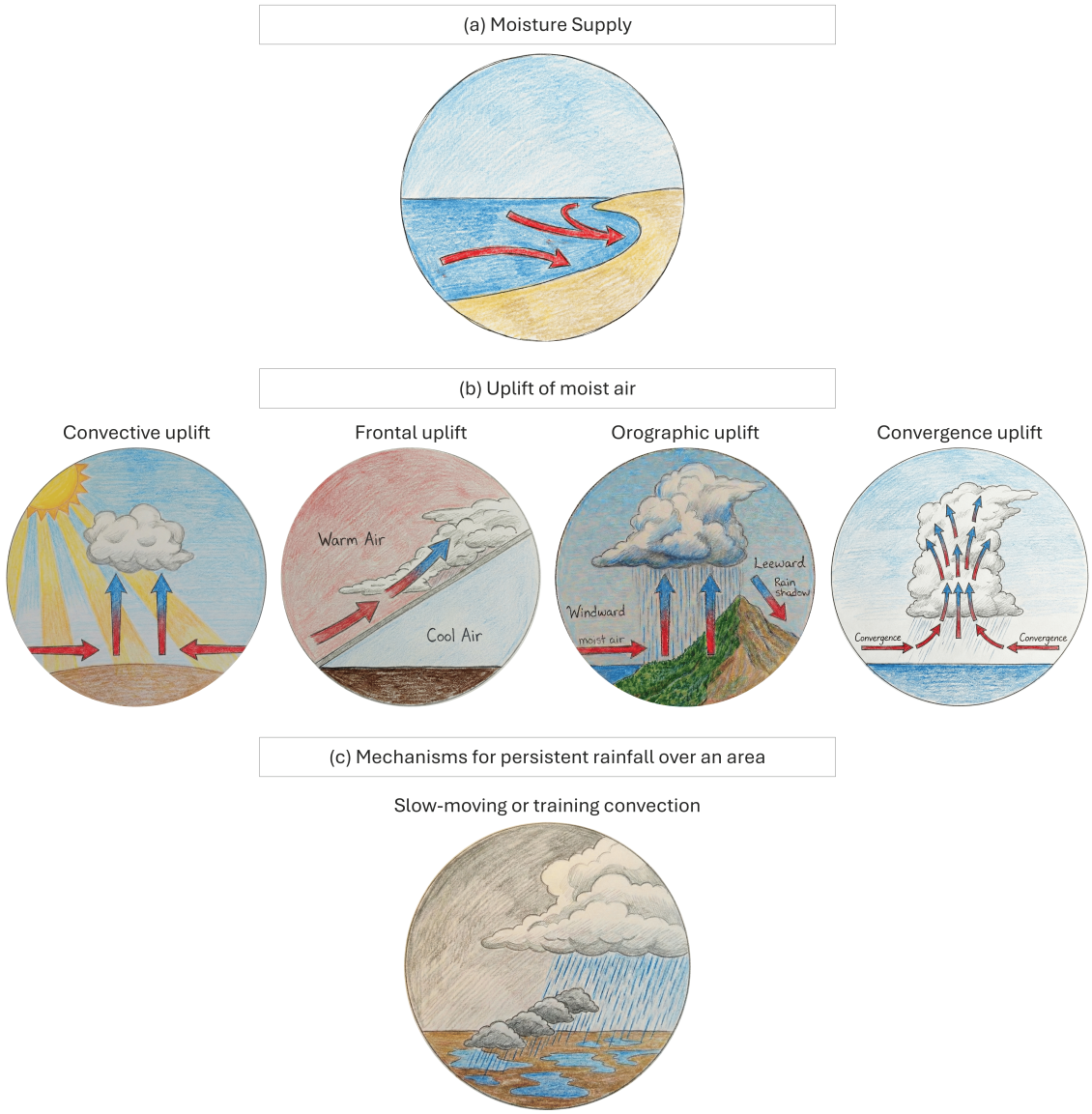

Figure 2.1: Key meteorological ingredients for flash flood-producing storms. Panel (a) shows the key ingredient number 1, i.e., the ample and persistent moisture supply. Panel (b) shows key ingredient n.2, i.e., the uplift of the moist air (convective, frontal, orographic, and convergence uplift). Panel (c) shows the key ingredient n.2, i.e., the mechanisms for persistent rainfall over an area.

also compels the moist air to rise. These mechanisms frequently operate in combination and vary in relative importance with geographic setting and synoptic conditions. The final ingredient relates to excessive precipitation when thunderstorms are slow-moving or stationary, continually reform over

the same area, or repeatedly traverse the same location through training convection or back-building (Figure 2.1c).

Since the 2000s, there have been numerous advances in the prediction of (localised) extreme rainfall with km-scale NWP models and ensembles, yet challenges persist

Inaccurate rainfall predictions have been identified as one of the most significant sources of uncertainty in flash flood prediction (Hapuarachchi et al., 2011; Zanchetta and Coulibaly, 2020). Since the early 2000s, there have been notable advancements in the short-range prediction (up to 2 days ahead) of (localised) extreme rainfall through the development of km-scale NWP models and ensemble approaches to quantify uncertainties (Roebber et al., 2004). Nonetheless, significant challenges persist. Clark et al. (2016) show how km-scale NWP models have been a step-change in our capabilities to predict (localised) extreme rainfall. However, to improve rainfall prediction of localised extreme events, in terms of both location and intensity, model resolution must extend beyond the grey-zone[2] and reach values of approximately 1 km (Castorina et al., 2022). Within the grey-zone, the partial representation of convective processes may lead to systematic errors in rainfall volume, which Sangati and Borga (2009) have identified as the primary source of error in flash flood predictions. Recent developments confirm this issue. For instance, Gascón et al. (2025) show that Destination Earth's deterministic model at 4.4 km resolution (within the grey-zone) provides similar skill to the 9-km ECMWF ENS for the Valencia flash flood in 2024. Hewson (2024) obtained similar results over an objective verification analysis over a more extended period of time. Although the experimental Destination Earth's 2.8 km run outperforms the 4.4 km forecasts, the probabilistic 9 km ensemble remains competitive. Moreover, this result suggests that higher resolution alone is insufficient and that gains in spatial detail must be accompanied by probabilistic representation to yield meaningful improvements in skill. Neighbourhood post-processing offers one approach to achieving this, generating a pseudo-ensemble by aggregating forecasts from adjacent grid-boxes to sample forecast uncertainty (Roberts et al., 2023). Finally, at medium-range lead times, flash-flood-producing convective storms exhibit very limited predictability, with forecast skill dropping sharply beyond a few hours (Buizza and Leutbecher, 2015;

[2]The *grey-zone* refers to NWP models with a grid-box resolution between 1 and 10 km (Wyngaard, 2004). At these grid-box resolutions, deep convective processes transition from being entirely parametrised (i.e., sub-grid physics) to being partly resolved. In the grey-zone, neither the parametrisation scheme nor the resolved dynamics alone provides a consistent representation of deep convective processes. Instead, both give a non-negligible contribution.

Barrett et al., 2019). In this regard, Jasper et al. (2002) shows that the primary limitation of extending flash flood predictions' lead times to medium-range timescales remains the precipitation forecast uncertainty between days 5 and 10. Schwartz (2019) and Schwartz and Sobash (2019) show that increasing the number of members computed at lower spatial resolutions can help increase the forecasts' lead time while providing a more cost-effective solution than simply increasing the forecasts' spatial resolution. However, this approach offers only a partial solution. Ensemble forecast spread can still be large, even at short-range lead times (Done et al., 2012), and important small-scale flash flood-triggering rainfall events may still be missed (Göber et al., 2008). These issues reflect the fact that the scope of the ensemble is to quantify uncertainties in rainfall prediction at the model's grid scale, ignoring what might be the rainfall at sub-grid scale (Bouallegue et al., 2020). Finally, while the development of 1-km rainfall forecasts would undoubtedly improve the prediction of areas at risk of flash floods, running a medium-range global NWP model at this resolution, even in deterministic mode, may remain computationally prohibitive for several decades (Zeman et al., 2021).

Global, lower-resolution ensemble NWP models offer skilful medium-range forecasts for extreme rainfall, with statistical post-processing enhancing skill for localised extremes

Global, lower-resolution (10+ km spatial resolution) ensemble NWP models, such as ECMWF's ENS at 9 km spatial resolution, could solve the issues mentioned above. They offer continuous global coverage and produce forecasts up to day 15. Moreover, the performance of such models has been increasing steadily, one day of lead time every decade (Bauer et al., 2015). However, for a long time, they have been considered inappropriate for flash flood prediction due to their coarse resolution, which cannot explicitly resolve small-scale intense convection (Schumacher, 2017). Low resolution notwithstanding, the interest in using rainfall forecasts from these ensemble global NWP models has been growing due to their increased performance, especially at medium-range lead times (Tripathy et al., 2020; Bucherie et al., 2022b). Additionally, statistical post-processing offers the possibility to successfully downscale the forecasts and provide skilful medium-range predictions for localised extreme rainfall (Vannitsem et al., 2021). One notable example is the post-processing approach called ecPoint, which post-processes global model outputs to provide probabilistic rainfall forecasts at point-scale (Hewson and Pillosu, 2021). ecPoint provides forecasts with better skill than raw ECMWF ENS (Hewson and Pillosu, 2021; Pillosu et al., 2025a) as well as raw and and post-processed km-scale rainfall forecasts (Gascón et al., 2024).

Barrier n.1: limited direct assessment of global NWP rainfall forecasts against flash flood occurrence

Despite there is a growing interest in rainfall forecasts from lower-resolution, global ensemble NWP models, they are not assessed explicitly for flash flood prediction. The assessment would be typically limited to rainfall observations, assuming that if rainfall was correctly predicted, the model would perform equally well when predicting flash floods (Gascón et al., 2024). However, while rainfall is the primary driver of flash floods, rainfall forecasts should be directly assessed against flash flood observations because there is not always a 1:1 correlation between the two events. There are two main ways to verify rainfall forecasts for flash flood prediction. The first compares rainfall forecasts directly against flash flood impact observations. This approach is typically qualitative and confined to local scales (Tripathy et al., 2020) due to the limited spatial extent of most impact databases (Gaume et al., 2009). Only a few studies consider large-scale domains such as the CONUS (Herman and Schumacher, 2016). The second approach uses rainfall forecasts as inputs to a distributed physics-based hydrological model to generate discharge predictions, which are then assessed against observed discharge using standard hydrological scores (e.g., KGE, NSE) to assess peak magnitude and timing (**?**). However, this method would face two main challenges if it were applied at global scale. First, there is no distributed physics-based hydrological model yet capable of producing discharge forecasts for small flashy catchments (i.e. below 100 km^2) at a global scale. Second, even if such forecasts were feasible, the severe paucity of discharge observations in these predominantly ungauged catchments (Clerc-Schwarzenbach et al., 2024) would preclude robust verification, including the disentanglement of compound hydro-meteorological uncertainties. Consequently, despite the potential of global NWP rainfall forecasts to extend flash flood prediction beyond data-rich regions, their suitability for this purpose remains largely unverified at the scales required for global application.

Opportunity n.1: flash flood impact databases can enable direct verification of rainfall forecasts against observed flash flood occurrence using probabilistic verification scores

The recent development of flash flood impact databases at global scale (e.g. EMDAT and DesInventar (Panwar and Sen, 2020)), continental scale (e.g., in the US (see Section 4.2) or in Europe (Dotzek et al., 2009), and national scale (e.g., in Ecuador (Bucherie et al., 2022a))) have opened up to the possibility of using impact flash flood data as a proxy observation for the identification of areas at risk of flash flooding. They consist of tabular data, in which each data point indicates the location (in the form of latitude/longitude coordinates) and the reporting date-time of a flash flood event. If the data points in such databases are pre-processed to indicate in which NWP grid-boxes flash floods were observed (binary yes- or no-events), they could be

used to run a first assessment on whether the global NWP rainfall forecasts identify areas at risk of flash floods. Despite of not being possible to assess the timing, location, and magnitude of the flash flood event with scores commonly used in hydrology because the rainfall is not transformed to discharge, it would still be possible to assess whether the rainfall forecasts can provide guidance on what areas are likely to experience flash flooding by adopting commonly used scores in the probabilistic verification of weather forecasts, such as reliability diagrams, frequency bias, and ROC curves (Jolliffe and Stephenson, 2012; Wilks, 2020). While this verification provides an outlook of flash flood risk, it proves useful for understanding whether the rainfall forecasts contain any useful signal regarding potential flash flooding (Mason, 1979). Two aspects must be considered when implementing any verification framework using impact flash flood reports. First, this type of observations are not static as measurements from rain or discharge gauges, which report yes-flash-flood-events (from now on, yes-events) and no-flash-flood-events (from now on, non-events). With the latter, it is indeed possible to assess whether the model produces a correct prediction (in the case the event was or was not predicted, and the event did or did not happen) or a wrong prediction (in the case the event was or was not predicted, and the event did not or did happen). Impact databases include only events that were reported, and it is not known whether a non-event represents an event that occurred but was not reported, or it represents an event that did not happen (Marsigli et al., 2021). Moreover, we do know that impact databases severely underestimate the frequency of flash flood occurrence (Panwar and Sen, 2020). This uncertainty makes it difficult to assess whether the prediction of a yes-event was correct or not (Robbins and Titley, 2018). Hence, it is likely that any verification would overestimate the cases in which correct forecasts might be considered as wrong, undermining the assessment of the forecasts and their potential usability in the prediction of areas at risk of flash flooding (Marsigli et al., 2021; Robbins and Titley, 2018). Despite these difficulties, the verification of rainfall forecast against impact flash flood reports may shed light on the capabilities of global, lower-resolution ensemble NWP models in identifying areas at risk of flash floods.

The research developed on this topic is presented in Chapter 5.

2.2 Are medium-range, data-driven hydro-meteorological predictions of areas at risk of flash floods feasible using reanalysis, forecasts, and impact flash flood reports with low spatial resolution?

Catchment characteristics and streamflow dynamics influence the catchment response to heavy rainfall and flash flood occurrence

Catchment characteristics (e.g. morphology, topography, land use, and soil properties) play a crucial role in determining the hydrological response of the catchment to heavy rainfall (Henao Salgado and Zambrano Nájera, 2022; Duan et al., 2022). The scale and size of catchments add complexity to hydrological responses, with larger catchments exhibiting greater heterogeneity in their hydrodynamic behaviours (Luong et al., 2021). Catchment topography significantly modifies water movement patterns, with steeper slopes promoting quicker runoff travel times and concentrated flow pathways (Maqtan et al., 2022). Antecedent soil moisture and infiltration rates (determined based on soil type and vegetation cover) also influence runoff generation through their effect on infiltration capacities and connectivity between surface and subsurface flow pathways (Zhai et al., 2018). Flash flood forecasting systems, indeed, rely heavily on accurate initial soil moisture estimates to simulate the catchment's hydrological response to rainfall (Yatheendradas et al., 2008).

The historical foundation for the understanding of flash flood occurrence in the second half of the 20th century: early methods offered limited modelling capabilities

The foundations of flash flood forecasting trace back to the second half of the 20th century and were rooted in simple, well-established hydrological modelling approaches (Beven, 2025): conceptual rainfall-runoff models (e.g., unit hydrograph, storage-based models), empirical or statistical approaches (e.g., regression-based relationships between rainfall and runoff), and, in particular, physics-based process models. These models employ principles derived from physical laws and equations to simulate hydrological processes and capture the intrinsic behaviour of natural processes such as precipitation-runoff transformations, soil infiltration, and river routing without relying on statistical or empirical relationships (Singh et al., 2024; Panigrahi et al., 2025). One of the earliest examples of a flash flood forecasting system, developed in the US (Georgakakos, 1987), relied primarily on simple mathematical representations of watershed behaviour, with limited data collection capabilities and processing power. Moreover, they depended heavily on ground-based observations using in situ sensors, which provided relatively limited spatial coverage and temporal resolution.

The historical origins of operational flash flood forecasting systems in the late 20th century: systems were limited by data availability and computational constraints, but started integrating deterministic NWP rainfall forecasts to extend the lead times of the hydrological forecasts

The late 20th century saw increasing efforts to translate theoretical hydrological understanding into operational flash flood forecasting systems. This period saw growing recognition that flash floods constitute a distinct hydrological hazard from other flood types, requiring specialised prediction approaches. As computational capabilities improved in the 1990s, more sophisticated hydrological models began to emerge. While still limited by data availability and computational constraints, flash flood forecasting systems started integrating rainfall predictions from NWP models (Collier, 2007): while deterministic, the integration of rainfall forecasts allowed for extending the forecast lead times from minutes to hours.

The birth of modern probabilistic operational flood forecasting systems in the early 21st century: integration of diverse hydro-meteorological data sources and advanced modelling capabilities due to increased computational resources, but still at catchment or national level

The beginning of the 21st century marked a pivotal moment in flash flood forecasting. The ability to integrate diverse data sources was key in the rise of modern flash flood prediction, overcoming the limitations of any single data type and providing more robust precipitation inputs to the hydrological models (Singh and Woolhiser, 2002; Todini, 2011). Such sources included remotely-sensed precipitation from satellite and radar, ground-based gauge measurements, and probabilistic rainfall forecasts from NWP models (Gouweleeuw et al., 2005). Moreover, the computational ability to process vast quantities of data in near-real time allowed forecasting systems to create timely predictions for flashy catchments (Liang et al., 2016; Leong et al., 2017). Despite these technological advances, the developed forecasting systems remained confined to small-scale domains, such as urban (Speight et al., 2018; Ibarreche et al., 2020) or national (Javelle et al., 2016). Such limitations were primarily due to the dependence on dense observational networks only available in geographically limited areas (Gaume et al., 2009) and on hydrological models that could not be upscaled to larger domains given the computational constraints of the early 2000s (Hapuarachchi et al., 2011).

Continental flash flood forecasting systems emerged from the beginning of the 2010s, yet their computationally intensive nature confined early developments to data-rich regions, leaving much of the Global South under-covered

From the beginning of the 2010s, prototypes of continental flash flood forecasting systems started to emerge. Several key technological breakthroughs enabled this extension (Philipp et al., 2016; Zanchetta and Coulibaly, 2020): improved predictions of extreme rainfall from global NWP models, the introduction of ensemble prediction systems that quantify forecast uncertainty, high-performance computing allowing more sophisticated modelling, improved precipitation measurements through satellite and radar products, and the improved efficiency of distributed hydrological models upscale to larger-scale domains. A notable example of those improvements is the European Flood Awareness System (Thielen et al., 2009), which,

despite not being designed specifically for flash floods, started producing (flash) flood predictions in large- to medium-sized (between 500 and 1000 km^2) catchments in Europe up to medium-range lead times (Bartholmes et al., 2009). In the following years, EFAS also incorporated two indices predicting areas at risk of flash floods: EPIC, based on rainfall (Alfieri and Thielen, 2015), and ERIC, based on soil moisture and rainfall (Raynaud et al., 2015). More recent developments improved discharge predictions of flash floods in medium- to small-sized catchments (between 100 and 500 km^2), where hydrological and routing processes are equally important as the performance of the NWP rainfall forecasts (Mazzetti et al., 2021). Other notable examples of large-scale operational flash flood forecasting systems predicting discharge in catchments between 1000 and 100 km^2 can be found primarily in the US (Clark et al., 2014; Gourley et al., 2017; Georgakakos et al., 2022), including also an ensemble configuration such as the EF5 distributed hydrological model (Flamig et al., 2020). More recently, China has also been at the forefront of model development for the prediction of flash flood events, but primarily at catchment/regional level (Liu et al., 2018). A common feature of these systems is the need to process vast amounts of hydro-meteorological data, volumes that grow exponentially as target catchment size decreases. Consequently, the high spatial and temporal resolution required to capture flash flood processes makes operational systems based on physics-based models extremely computationally intensive (Efstratiadis and Koutsoyiannis, 2010). For this reason, early developments of large-domain flash flood forecasting systems occurred in economically developed and data-rich regions, such as the US, Europe, and China, leaving many regions of the Global South under-covered or unprotected (Nearing et al., 2021).

The rise of data-driven models for flash flood prediction since the late 2010s to overcome the limitation of physics-based flash flood predictive models

Between the late 2010s and the early 2020s, data-driven models have gained prominence in flash flood prediction with the aim of overcoming the previously discussed limitations of traditional physics-based flash flood models (Mosavi et al., 2018; Al-Rawas et al., 2024; Byaruhanga et al., 2024). Opposite to physics-based hydrological models, data-driven approaches learn rainfall-runoff relationships directly from empirical data, bypassing the need for explicit process parametrisation (Chang and Chen, 2018; Zhang et al., 2018). This empirical foundation also enables data-driven models to capture complex patterns in flash flood generation that remain incompletely understood from a theoretical standpoint (Sun and Scanlon, 2019). A recent systematic review by Santos et al. (2025) indicates that data-driven techniques feature in the majority of new flash flood prediction studies,

reflecting a substantive shift toward data-driven modelling. Among the most commonly employed architectures, Long Short-Term Memory (LSTM) networks dominate, appearing in 60% of the reviewed studies. Multilayer Perceptrons (MLP, 28%), Convolutional Neural Networks (CNN, 16%), and Support Vector Machines (SVM, 14%) followed as the most commonly used data-driven architectures (Santos et al., 2025). Despite LSTM prevalence, no single architecture consistently outperforms others across all examined models, suggesting that model performance remains highly dependent on regional hydrological characteristics, input data quality, and forecast horizon (Santos et al., 2025). Data-driven models for the prediction of flash floods typically ingest rainfall observations or forecasts as primary inputs (used in 88% of studies), and consider discharge (46%) or water level (38%) measurements as target variables (Santos et al., 2025). Other key considerations in the development of data-driven flash flood prediction models include the selection of predictive features, the quality of historical flash flood observations, suitable model architectures, uncertainty quantification, and the computational trade-offs inherent in operationalising such systems (Santos et al., 2025). From an operational standpoint, data-driven models also offer significant advantages when compared to physics-based models: although training may demand substantial computational resources, inference is rapid, typically generating predictions within seconds to minutes, which is well-suited to the real-time demands of rapid onset flash floods (Guo et al., 2021; Liao et al., 2023). Moreover, these models can be efficiently retrained as new data become available without requiring expensive model recalibrations (Zhou et al., 2025).

Barrier n.2: most data-driven studies rely on high-resolution inputs (km-scale rainfall forecasts) and conventional target variables (e.g., discharge or water levels): this leaves unexplored the possibility to create data-driven predictions of areas at risk of flash flood with low-resolution NWP forecasts and impact reports

As shown in the previous paragraph, despite their promise, most data-driven flash flood prediction studies rely on high-resolution inputs (e.g., rain gauges, weather radar, or km-scale NWP forecasts) and dense discharge or water level measurements as target variables (Santos et al., 2025). This configuration constrains applications to limited data-rich domains and short-range lead times. Creating a data-driven flash flood prediction model that produces medium-range forecasts over a continuous global domain requires addressing two fundamental questions. First, can lower-resolution but globally available inputs, such as reanalysis products or ensemble NWP forecasts, provide sufficient signal for flash flood prediction? As discussed in Section 2.1, verification of global NWP rainfall forecasts against flash flood impact observations suggests that useful predictive guidance may be retained despite the inability of coarse NWP models to resolve fine-scale convective processes (Pillosu et al., 2024). Second, and largely unexplored,

is whether flash flood impact reports can serve as viable target variables for supervised learning in the prediction of areas at risk of flash flood. Such reports offer broader geographic coverage than discharge and water level observations but present several challenges (Panwar and Sen, 2020; Marsigli et al., 2021; Pillosu et al., 2024): they are inherently binary, indicating only yes- and non-events without magnitude or severity information; they report only observed events, making it impossible to distinguish true non-events from unreported occurrences and risking an overestimation of false alarms; they are spatially imprecise, often reported at administrative unit level rather than at catchment scale; and finally, they are temporally uncertain, as the reporting time may not correspond precisely to the real time at which the event occurred. These characteristics contrast sharply with the continuous, georeferenced discharge or water level measurements that conventional data-driven hydrological models are designed to predict.

Opportunity n.2: lower-resolution global forecasts and binary impact reports may enable medium-range flash flood predictions over a continuous global domain (regional development and verification)

To date, no study has systematically investigated whether the combination of low-resolution global forecasts and binary impact reports can enable medium-range flash flood predictions over a continuous global domain. However, analogous problems have been addressed in other fields: Cavaiola et al. (2024) shows the benefits of data-driven approaches for lightning, a problem exhibiting similar class imbalance challenges, achieving skilful forecasts up to t+48. This precedent suggests that, with appropriate model architectures and training strategies to address class imbalance, data-driven flash flood prediction using lower-resolution inputs and binary targets may be feasible. Furthermore, whilst the limitations of impact reports cannot be denied, they nonetheless present two important advantages over conventional discharge or water level observations. First, although impact reports preclude numerical predictions of discharge or water level, they enable binary (i.e., yes- or non-event) predictions of areas at risk of flash flood. This is precisely the operational information that forecasters typically provide to users, especially at longer lead times where compounding uncertainties makes more precise numerical predictions unsuitable for decision-making (Zanchetta and Coulibaly, 2020). Second, while a high-density global flash flood impact database does not yet exist, the potential for a rapid expansion is considerably greater than for conventional hydrological measurements. Impact reports can today be gathered through internet-based sources at minimal cost (Wyatt et al., 2023, 2024), whereas extending and maintaining discharge or water level monitoring infrastructure may remain prohibitively expensive with little prospect of substantial expansion in the next decade (McCabe et al., 2017; Andrews and Grantham, 2024; Nasta et al., 2025).

Hence, investigating whether impact reports can successfully be used as target variables in data-driven approaches to generate skilful predictions of areas at risk of flash floods represents a necessary step toward extending flash flood forecasting beyond data-rich regions and short-range timescales. However, testing this hypothesis requires at least a continental domain with a high-density observational dataset (such as the US) where model development and verification can be conducted with confidence. Regions such as the United States, which maintain a comprehensive flash flood impact database, offer the necessary foundation for evaluating whether this approach yields skilful predictions before any transfer to less-monitored areas is attempted.

The research developed on this topic is presented in Chapter 6.

2.3 How does the coverage-density trade-off influence training data strategies to develop predictions of areas at risk of flash floods over a continuous global domain

Data-driven models trained on diverse catchments generalise beyond their training domain, challenging the assumption that global predictions require uniformly distributed high-density observations

Traditional understanding holds that developing hydrological predictions over a continuous global domain requires high-density observations uniformly distributed across the globe, as regional differences in rainfall climatology and hydrological characteristics demand locally representative training data (Alfieri et al., 2013; Hrachowitz et al., 2013). Without such data, flood predictions in ungauged regions can exhibit substantial errors (Prakash et al., 2025). At most, regionalisation techniques were applied to develop hydrological prediction in ungauged catchments adjacent to gauged catchments (He et al., 2011). As discussed in section 2.2, this assumption also shaped early developments of data-driven hydrological modelling. Especially for deep architectures, data-driven models for the prediction of flash floods offer robust performance only where observational data is abundant and of good quality (Zanchetta et al., 2022; Gacu et al., 2025). This characteristic has confined applications to data-rich regions and primarily at catchment/regional scales, leaving the Global South largely without flash flood forecasting capability (Nearing et al., 2021). More recent developments have demonstrated that data-driven hydrological models, trained on standardised large-scale hydrological datasets with comprehensive catchment attributes and records for forcing (i.e., rainfall forecasts) and target

variables (e.g. discharge), can achieve remarkable performance within the training domain. For example, a model based on a Long Short-Term Memory (LSTM) neural network for the CONUS, leveraging CAMELS US (Addor et al., 2017), outperformed traditional hydrological models (Kratzert et al., 2018). Similarly, regional flash flood prediction systems in other continents have capitalised on dense networks[3] of weather radars and stream gauges to develop high-performing localised models (Santos et al., 2025). These success stories underscore the value of data density in capturing the complex, non-linear relationships between meteorological forcing and flash flood occurrence. However, a key insight from this work was that these data-driven models were also able to generalise to ungauged catchments within their training domain when trained on sufficiently diverse hydro-climatological conditions (Kratzert et al., 2019). This finding raised a compelling question: if diversity within a region enables generalisation, could combining observations from multiple regions extend this capability across continents or even globally? This hypothesis has prompted a growing movement to harmonise CAMELS-style datasets across regions and continents, facilitating large-sample hydrology at increasingly broader scales (Clerc-Schwarzenbach et al., 2024). This made the concept of hydrological similarity evolve from simple geographical proximity to sophisticated multi-dimensional feature spaces encompassing climate, topography, geology, and land use characteristics (Addor et al., 2020). This approach has indeed been demonstrated to be successful for the prediction of riverine floods (Nearing et al., 2024) but also in the prediction of areas at risk of unseen or extremely severe flooding (Bertola et al., 2023).

Barrier n.2: large-sample hydrology teaches us that global data-driven models for riverine flood prediction can be achieved with currently available low-resolution hydrological observations around the world, yet flash flood observations are orders of magnitude more scarce and unevenly distributed

Flash floods are relatively infrequent and highly localised events that often occur in ungauged basins lacking discharge or water level records (Gaume et al., 2009). This results in a severe paucity of labelled target data for supervised learning, several orders of magnitude greater than for riverine floods. To clarify, in the CARAVAN dataset, only 0.1% of the catchments have a size < 100 km^2, leaving only a handful of data points suitable for training a data-driven model for the prediction of flash floods. Moreover, many events are documented only anecdotally through post-event reports, disaster databases, or news media. While some regions have assembled

[3]Germany is also added in this list as it has started a project led by scientists at the Karlsruhe Institute of Technology, to provide nationwide data-driven predictions for small catchments (< 500 km^2).

catalogues of flash flood impacts, e.g., the US (see Chapter 4.2) and Europe (Dotzek et al., 2009), there is no unified global archive with consistent reporting criteria for flash floods. Moreover, many events in remote regions may go unrecorded, introducing under-reporting bias (Panwar and Sen, 2020; Marsigli et al., 2021). Can data-driven approaches for flash flood prediction work with such a small training dataset? In particular, the limited sample of flash flood occurrences creates a severe class imbalance, with non-events vastly outnumbering the recorded yes-events. When training datasets are dominated by non-events, models may default to predicting non-events as the statistically optimal response, effectively failing to capture the rare events of interest (Kaur et al., 2019). Other approaches have considered the use of long timeseries of modelled hydrological data, e.g. global hydrological reanalysis (Harrigan et al., 2020) as a proxy for observations in hydrological modelling over large domains (Prudhomme et al., 2024). Whilst in theory this approach allows us to train cheap data-driven hydrological models over large continuous domains, and could potentially also be used in the prediction of areas at risk of flash floods, the quality gap between direct observations and proxy data sources remains a significant problem that still makes us strive for training data-driven models with actual observations. For example, Zsoter et al. (2019) found that ERA5 systematically underestimates extreme precipitation intensities by 20-40%.

Opportunity n.3: leveraging a data-rich region with diverse hydro-climatological conditions to train data-driven flash flood models and evaluate their transferability to data-sparse areas

Kratzert et al. (2024) have theorised that it is significantly better to train one single data-driven hydrological model with data from wherever it is available. This thesis pursues an analogous approach for flash flood prediction. However, the extreme paucity of flash flood impact reports compared to the discharge observations available for riverine flood modelling necessitates a more nuanced exploration of training strategies. This thesis adopts the balanced approach discussed by Gupta et al. (2014) between depth and breadth. Rather than focus the learning on individual catchments (depth), the aim is to leverage a data-rich region, such as the US, with sufficient hydro-climatological diversity (breadth) to discover general principles of flash flood occurrence that may transfer beyond the training domain. In particular, a comprehensive sensitivity analysis will be carried out to determine the optimal balance between spatial coverage and data density when training a data-driven model to create predictions of areas at risk of flash floods over a continuous global domain. In fact, a high-quality and high-density flash flood impact database in a varied hydro-climatological region, such as the Storm Events Database in the US, offers an invaluable

opportunity to evaluate whether models trained in data-rich regions can successfully transfer to data-sparse areas.

The research developed on this topic is presented in Chapter 7.

CHAPTER 3

INTEGRATED EXPERIMENTAL STRATEGY

This chapter outlines the integrated experimental strategy adopted in this thesis, wherein each research question and objective - presented in Chapter 1 - builds upon its predecessor to establish a proof-of-concept system for medium-range predictions of areas at risk of flash floods across a continuous global domain (upper panel in Figure 3.1). The integrated experimental strategy is here exemplified through the *methodological decisions* underlying each research question and objective (lower panel in Figure 3.1). Such decisions encompass three primary areas: the selection of appropriate data sources (Section 3.1), the strategy for developing a data-driven model to identify areas at risk of flash floods under imbalanced observational datasets (Section 3.2), and the formulation of the forecast verification strategy (Section 3.3).

3.1 Data requirements

The development of a robust flash flood prediction system necessitates a rigorous, structured assessment of data requirements, as the quality of the input features directly constrains the forecasts' accuracy and reliability. This selection process must address distinct needs across the entire methodological lifecycle, i.e., *model development*,*forecast verification*, and *future operational implementation*. This section outlines the specific physical and operational criteria established to select the observational and forecast datasets that underpin this study.

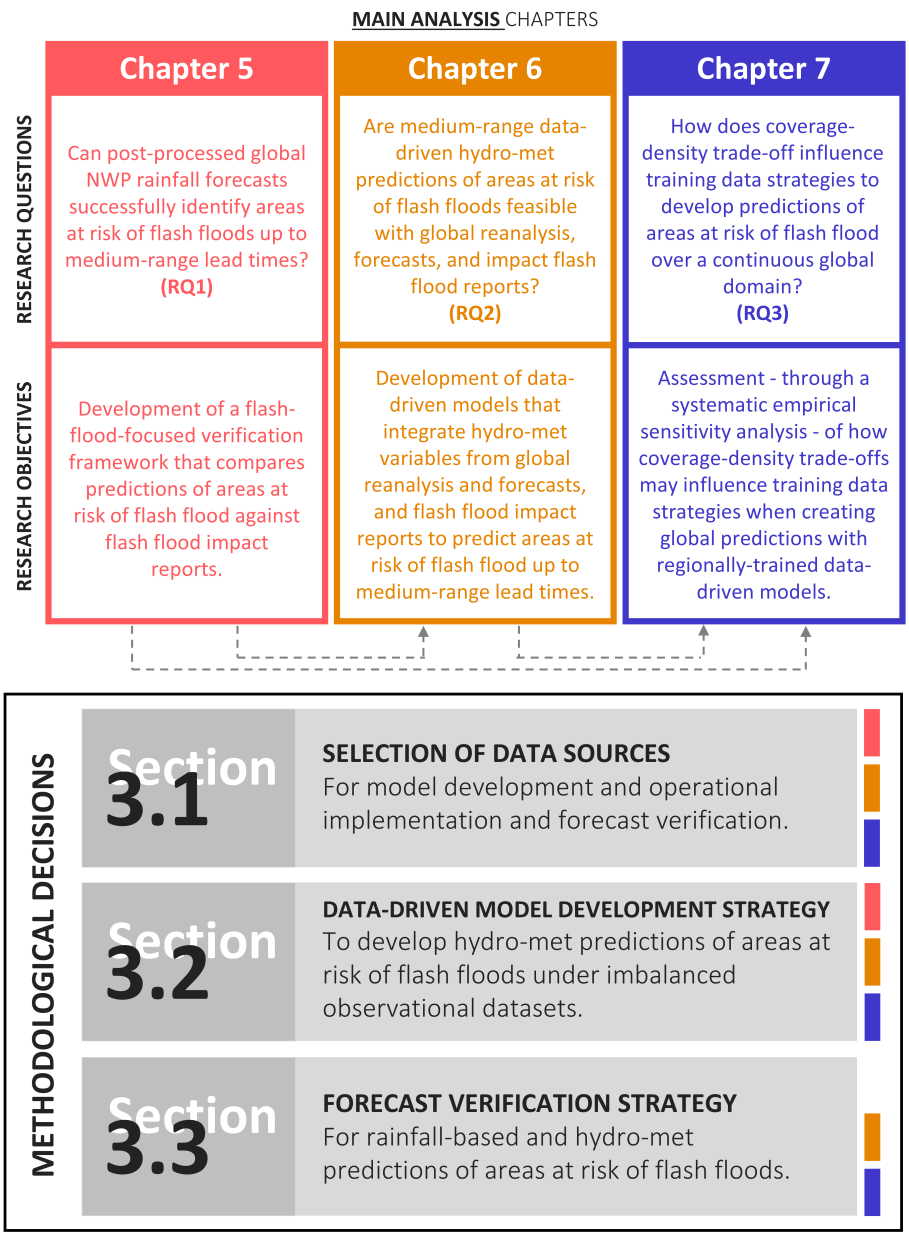

Figure 3.1: Thesis' integrated experimental strategy. The upper panel of the infographic reminds the reader about the hierarchical relationship between research questions RQ1 (addressed in the Main Analysis Chapter 5, card in pink), RQ2 (Main Analysis Chapter 6, card in yellow), and RQ3 (Main Analysis Chapter 7, card in blue) and corresponding research objectives. The horizontal dashed arrows beneath each card indicate the cross-chapter information flow as introduced in Chapter 1. The lower panel of the infographic (within the solid black box) identifies the three core methodological decisions that inform the integrated experimental strategy: data source selection (Section 3.1), forecast verification strategy (Section 3.3), and data-driven model development strategy (Section 3.2). The coloured indicators on the right of each grey card identify the main analysis chapter in which each methodological decision was applied.

3.1.1 Requirements for observational data

Requirement for observational data n.1: need to represent both types of flash floods, fluvial and pluvial

Two categories of observational data can be used in this thesis to predict areas at risk of flash floods: impact-based reports and gauge-based measurements. Impact-based reports are chosen over gauge-based measurements as the objective of this thesis is to predict areas at risk of both fluvial flash floods - occurring within river channels - and pluvial flash floods - resulting from inadequate urban drainage systems or surface runoff accumulation, including areas away from river channels. Discharge gauges would capture only fluvial flash floods. Furthermore, gauge-based observations systematically underestimate flash flood frequency due to a sparse observational network and the tendency for flashy catchments to remain ungauged (Gaume et al., 2009, 2016). Although impact databases are subject to reporting biases — notably those related to population density (Marjerison et al., 2016) — they provide the most feasible approach for developing and verifying continental-scale predictions of areas at risk of flash floods.

Requirement for observational data n.2: spatio-temporal accuracy of reported flash flood events

Given that flash floods are rapid-onset, localised events, typically occurring in small catchments, observational datasets must accurately capture both the location and timing of each event, alongside quantifying the spatio-temporal uncertainty associated with each record. This precision is essential for developing a robust flash-flood-focused verification framework (RQ1), where grid-based assessments require accurate spatial assignment of events to model grid cells and accurate temporal assignment to accumulation periods. Furthermore, precise spatio-temporal information enables the establishment of meaningful relationships between local meteorological conditions and flash flood occurrence, thereby capturing the fine-scale processes that govern these rapid-onset events (RQ2 and RQ3).

Requirement for observational data n.3: long, complete, and consistent timeseries of reported events.

The rarity of flash flood events necessitates extensive temporal coverage to accumulate an adequate number of events for robust forecast verification (RQ1 to RQ3) and data-driven model training (RQ2). Additionally, consistent reporting standards across the considered spatial domain ensure a less biased verification and prevent the introduction of artificial patterns during model training. Moreover, consistent reporting standards would facilitate sensitivity analyses for expanding regional training to global scales, as suitable global databases are not available (RQ3).

This thesis employs flash flood impact reports from NOAA's Storm Events Database over the CONUS as the primary observational data

source. For a more detailed description of this dataset, please refer to Section 4.2 in Chapter 4.

3.1.2 Requirements for hydro-meteorological reanalysis and forecast data

Requirement for hydro-meteorological forecasts n.1: minimal disruption in the spatial resolution of reanalysis and forecast dataset when transitioning from training to inference stage

To generate a seamless transition from training (using reanalysis datasets) to inference stage (using forecast datasets), it would be preferable to use a system that provides a consistent spatial resolution across the training and the forecasts datasets, specially at this ”proof-of-concept” stage of model development. Varying resolutions would compound extrinsic uncertainties with those intrinsic ones, arising from both the chaotic nature of flash-flood-generating rainfall events and the complexity of the hydrological processes involved in flash flood generation. By employing consistent spatial resolutions throughout, any training-to-forecast and lead-time-driven degradation in model performance is more likely to reflect genuine predictability limits rather than dataset inconsistencies.

Requirement for hydro-meteorological forecasts n.2: Forecasts with continue global coverage

Although the proof-of-concept presented in this thesis focuses on the creation (RQ2 and RQ3) and verification (RQ1 to RQ3) of predictions of areas at risk of flash floods over the CONUS (i.e., a continental-scale domain selected specifically due to the high-quality observational data that covers it), the underlying methodology is explicitly designed to support global scalability. This imposes a strict constraint on the training and forecasts datasets used to define the model’s input features: they must not rely on datasets with only regional coverage, such as radar or rain gauges. Instead, the requirement for global extension necessitates the use of spatially continuous and physically consistent datasets that cover the entire globe without regional discontinuities. Consequently, this study relies exclusively on global NWP model outputs, which provide uniform hydro-meteorological forcings worldwide, ensuring the system can be seamlessly transferred to other regions without structural modification.

Requirement for hydro-meteorological forecasts n.3: capability to represent flash-flood-triggering rainfall events

Due to their typical coarse resolution (> 10 km) and parametrisation schemes of convective systems, raw global NWP model outputs tend to systematically underestimate localised rainfall extremes, which are critical for flash flood generation. Localised rainfall peaks tend to be underestimated in the case of large-scale rainfall (whether from stratiform rainfall or large convective systems such as extratropical and tropical cyclones) or absent in the case of isolated, storm-scale convection. Consequently, both

the verification of areas at risk of flash floods (RQ1 to RQ3) and the training of machine learning models for predicting such areas (RQ2) require rainfall estimates that can identify flash-flood-triggering rainfall events.

This thesis uses hydrological and static parameters from ERA5 reanalysis and forecasts because they fulfil requirements 1 and 2, while ERA5-ecPoint post-processed rainfall reanalysis and forecasts fulfil requirement 3. Please, refer to Sections, respectively, 4.3 and 4.4 in Chapter 4 for a more detailed description of these datasets and the variables used in this thesis.

3.2 Requirements for the development of data-driven models trained on severely imbalanced datasets

Requirement n.1: quantify uncertainty in data-driven model outputs trained on severely imbalanced datasets to support actionable decision-making in operational contexts

Rigorous uncertainty quantification is essential to produce reliable and skilful forecasts for rare events such as flash floods. This need is even greater if the forecasts are created with a data-driven model trained on a severely imbalanced dataset. The training database used in this study (see Section 4.2) contains merely $\sim$0.2% of yes-events (i.e. when a flash flood event is reported). As a result, the data-driven model outputs could converge to trivial classifiers, achieving accuracies exceeding 99% by exclusively predicting non-events, which would have no operational use. This is a problem also found in lightning (Cavaiola et al., 2024) and landslide detection (Xu et al., 2022; Agrawal et al., 2017; Zhang et al., 2022; Gupta and Shukla, 2023). Hence, rather than producing a single deterministic output, the system must capture intrinsic prediction uncertainties associated, for example, with the detection of flash-flood-triggering rainfall events and the interaction of such rainfall with the catchments' hydrology. To achieve this, it is important to prioritise methods that provide probabilistic outputs or ensemble-based variability, ensuring that operational decision-makers can interpret not just where a flood might occur, but how reliable that prediction is given the constraints of the training data.

Requirement n.2: for a preliminary, "proof-of-concept" stage, prefer algorithm-level over data-level approaches to quantify the uncertainty in the data-driven predictions

At a preliminary, "proof-of-concept" stage, the need for uncertainty quantification to produce robust model outputs should be satisfied by the use of algorithm- and ensemble-level approaches rather than data-level methods. Algorithm-level approaches modify the learning process through techniques such as weighted learning functions, while ensemble-

level methods, such as ensemble-based algorithms or cross-validation, combine multiple classifiers or train many models over different samples of the training dataset to capture the inherent uncertainties in predictions generated with imbalanced observational datasets. Moreover, if multiple ensemble-based algorithms are chosen, one can also examine how distinct algorithms capture the subtle signals preceding flash flood events in the training data. In contrast, data-level approaches, also known as sampling methods and including oversampling, undersampling, and synthetic data generation, modify the training dataset. This alters the original ratio between yes- and no-events and can potentially obscure the true rarity of flash flood events, compromising uncertainty estimates. By preserving the original ratio in the observational dataset, the developed models provide probability estimates that are more likely to reflect actual flash flood occurrence patterns, and prediction uncertainties may remain more interpretable for operational decision-making. Whilst more sophisticated sampling methods may prove beneficial in future development stages, this initial proof-of-concept prioritises approaches that maintain data integrity, establishing a baseline against which more complex systems can be benchmarked.

For a more detailed description of the ensemble-level approaches considered in this study, please refer to Section **??** in Chapter 6.

3.3 Requirements for a robust verification of probabilistic predictions of areas at risk of flash floods

Requirement n.1: objective verification must assess both features of probabilistic forecasts - reliability and discrimination ability

The operational utility of probabilistic forecasts - in this case, probabilistic rainfall-based predictions of areas at risk of flash floods - relies on the concurrent high performance in two desirable properties: *reliability* and *discrimination ability* (Jolliffe and Stephenson, 2012; Wilks, 2020). Reliability measures whether, for a specific probability bin, the chosen verifying threshold is predicted with a probability that equals the average frequency at which such an event is observed within that probability bin. Discrimination measures the ability of the forecasts to distinguish between situations that lead to events exceeding the verifying threshold, at specific probability thresholds, and those that do not. As shown in Pillosu et al. (2025a) for rainfall, forecasts exhibiting high discrimination ability may nonetheless yield uncalibrated probabilities. Conversely, forecasts that achieve good

reliability but exhibit poor discrimination provide minimal operational utility as they cannot effectively identify when conditions deviate from climatological norms to a heightened event likelihood. Explicit examination of both features is needed through *objective verification*, i.e., through the use of appropriate statistical verification scores.

Requirement n.2: use of overall and breakdown scores to provide summaries of model performance as well as a granular performance assessment across discrete probability thresholds (especially useful in the assessment of extremes

Overall and *breakdown* scores must be considered to ensure a thorough evaluation of forecast reliability and discrimination ability for extremes as well as for average events. Overall scores integrate performance across all probability thresholds, providing valuable summaries for model comparison and evaluation. However, these aggregated metrics may mask critical performance variations at specific probability thresholds - relevant for the performance analysis of extreme events - through compensating errors across the whole probability spectrum. Moreover, operational applications may prioritise performance assessments at particular probability thresholds, providing a granular performance assessment across discrete probability thresholds and enabling a complete characterisation of forecast utility.

Requirement n.3: select metrics that assess how well the model identifies the minority, positive class (yes flash flood events) when forecasting low-frequency events

When evaluating predictions of binary outcomes (e.g., occurrence of yes- and non-events) for low-frequency events, the difference between the number of observed events falling into the ”yes-event” and ”non-event” classes may be large. In this case, one would say there is a *class imbalance* in the observational dataset, with the negative class typically outnumbering the positive one (which is the one of interest, e.g., flash flood occurrence). This class imbalance may render traditional accuracy metrics meaningless. For example, forecasts predicting exclusively non-events may achieve 99% accuracy. While such a model appears near-perfect, such a high accuracy value is overwhelmed by the high number of correctly identified non-events (Wilks, 2020). Consequently, the score reflects the baseline rarity of the event rather than assessing the model’s skill in predicting the minority, positive class (which is the one of interest).

Requirement n.4: carry out a subjective verification analysis for specific case-studies to complete objective verification results

In addition to analysing forecast performance through objective verification, it is argued that predictions should also be examined through *subjective verification* involving case studies. Provided there are enough observations to analyse a specific case study, this type of analysis may be more effective than objective verification in differentiating forecast performance and predictability in specific cases of interest. Objective verification might indeed yield statistically weak results if the observational sample size is small.

Section 5.2 in Chapter 5 describes in detail the verification scores considered in this thesis to address requirements 1 and 2: frequency bias and reliability diagrams, respectively, as overall and breakdown scores to assess forecasts' reliability, and ROC curves and area under the ROC curve, respectively, as overall and breakdown scores to assess forecasts' discrimination ability. These scores will also be applied to verify the data-driven forecasts of areas at risk of flash floods described in Chapters 6 and 7. Section 6.2 in Chapter 6 describes the verification scores chosen to assess whether the model forecasts are not giving up in the prediction of the minority class: the precision-recall curve (breakdown score) and the area under the precision-recall curve (overall score). These scores will also be applied to verify the data-driven forecasts in Chapter 7. Finally, the case study used throughout the thesis, Storm Ida, is described in detail in Section 4.5 in Chapter 4, while the discussions of forecasts' performance for this case study are shown in the three main analysis chapters (5 to 7).

CHAPTER 4

DATASETS

The Rationale for ERA5 reanalysis and forecasts, ecPoint post-processing, and NOAA Storm Event Database selection development of (data-driven) models for the prediction of areas at risk of flash floods requires considering multiple hydro-meteorological and static parameters that influence flash flood occurrence. This thesis employs the state-of-the-art ERA5 global reanalysis, which provides spatially and temporally consistent reconstructions of atmospheric and land surface conditions, essential for model training. ERA5 forecasts are used at the inference stage. Whilst the spatial resolution of ERA5 data - both reanalysis and forecasts - is low for flash flood applications ($\sim$31 km at the equator), the use of the same NWP model throughout training and forecast production benefits the analysis in this thesis by eliminating uncertainties due to changes in forecasts' spatial resolution at the inference stage. Moreover, working with lower resolution data is more cost-effective when running multiple experiments, as done in this thesis. Only rainfall estimates - the primary driver of flash flood occurrence - are post-processed to enhance the spatial detail and accuracy of predictions, particularly for extreme localised events. The adopted post-processing technique in this thesis is ECMWF's ecPoint. Additionally, the selection of appropriate observational records of flash flood impacts is crucial for model development and validation. While global impact databases exist (e.g., EMDAT or DesInventar), regional databases with higher spatial density and more comprehensive coverage of smaller-scale flash flood events provide superior data quality. This thesis uses the flash flood impact reports from the Storm Event Database over the CONUS.

4.1 Study domain: the CONUS

This research establishes the CONUS as the primary study domain for the flash-flood-focused rainfall verification, data-driven model development, and sensitivity analysis for the global extension of the regionally-trained models (Figure 4.1).

CONUS domain
Orography at 1 km and location of the 25 most populated cities

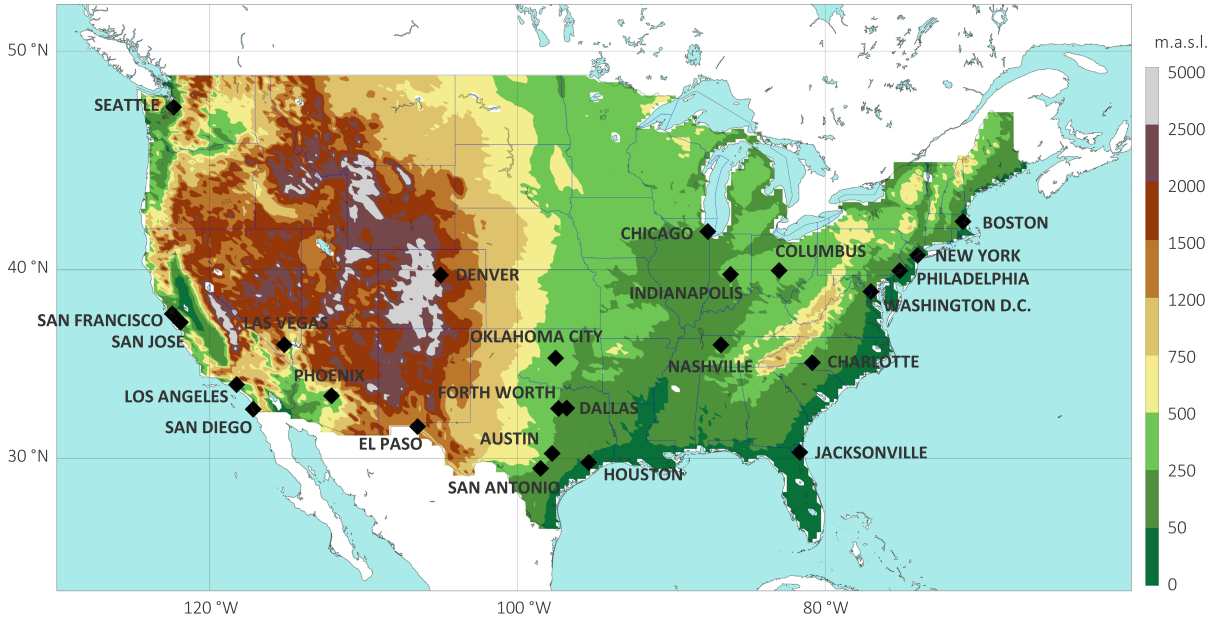

Figure 4.1: CONUS domain. The figure shows the orography at 1 km resolution (in shades of green and brown) and the location of the 25 most populated cities (black dots) over the CONUS.

CONUS as the primary study domain: justification

The selection of CONUS as the primary study domain is methodologically justified on several grounds. Foremost is the higher quality and comprehensiveness of the Storm Event Database maintained by the National Oceanic and Atmospheric Administration (NOAA). The high spatial and temporal resolution of the events reported, the consistent reporting protocols, and the rigorous quality control procedures adopted to populate this regional database mitigate the high data uncertainty and events underreporting that plagues global databases (Panwar and Sen, 2020; Gaume et al., 2009). A detailed description of the Storm Event Database is provided in Section 4.2. Moreover, the CONUS domain presents diverse hydrometeorological conditions, encompassing varied topographical features ranging from coastal plains to mountainous terrain, and climatic regimes spanning Mediterranean, continental, subtropical, and desert classifica-

tions. This diversity enables the development of a machine learning model exposed to a broad spectrum of flash-flood-generating mechanisms, from intense convective precipitation to rain-on-snow events, to hurricanes, and to large-scale systems such as atmospheric rivers, thereby enhancing its potential transferability to global applications (Dougherty and Rasmussen, 2019; Saharia et al., 2017).

4.2 Flash flood impact reports - NOAA's Storm Event Database

Storm Event Database: general description

NOAA's National Centers for Environmental Information (NCEI) Storm Event Database serves as the US's official repository of severe weather records. This comprehensive database has evolved significantly since its inception in 1950, transforming from solely county-level tornado reports to a sophisticated system documenting up to 48 different weather phenomena with increasingly precise geospatial representation[1]. Its historical progression affects data completeness for different phenomena. Flash flood records became standardised from 1996 onward, with major geographical precision improvements implemented in 2007 when reporting transitioned from county-level to a polygon-based representation[2]. The database covers the entire United States and territories with a spatial bounding box from 172.0°W to 65.0°E longitude and 18.0°N to 72.0°N latitude[3] (Figure 4.2a). Data collection involves a network of 123 NWS Forecast Offices gathering information from emergency management officials, law enforcement, trained SKYWARN spotters, damage surveys, media reports, and public observations[4]. This study uses version 3.1 of the database. It considers only reports over the CONUS that go from 2001 to 2024 (Figure 4.2b), as reports before 2001 do not contain latitude/longitude coordinates, which makes them unusable for model development and verification.

Storm Event Database: metadata quality

The Storm Event provides extensive metadata for flash flood events, compiled by professional meteorologists and hydrologists. Metadata

[1] https://www.ncdc.noaa.gov/stormevents/details.jsp

[2] https://inside.nssl.noaa.gov/flash/database/

[3] https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00510

[4] https://www.ncdc.noaa.gov/stormevents/faq.jsp

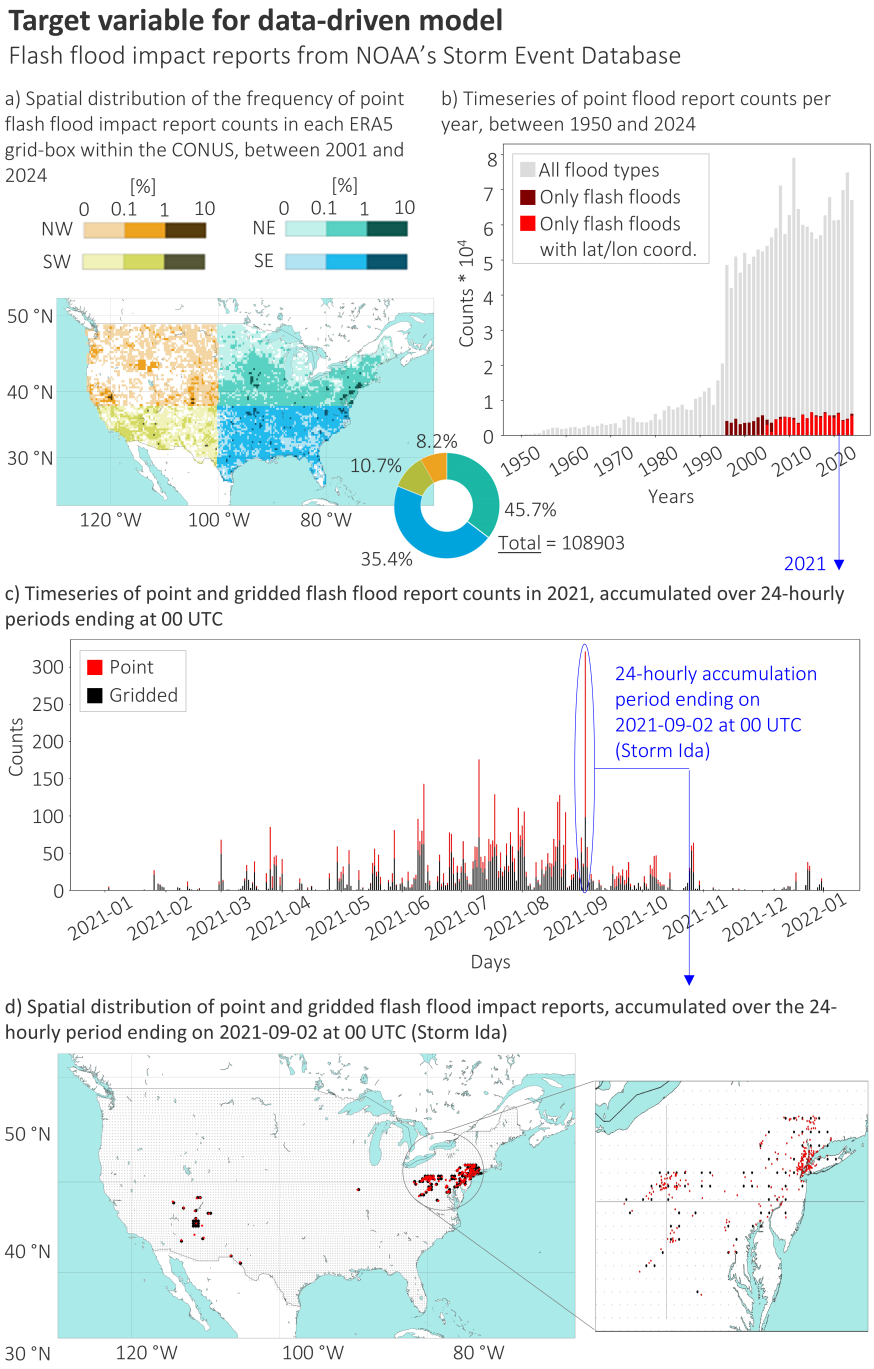

Figure 4.2: Flash flood reports in NOAA's Storm Event Database. Panel (a) shows the point flash flood report frequencies for each grid-box within the CONUS, between 2001-2024. Four quadrants shown: North-West (NW, orange shades), North-East (NE, green), South-East (SE, blue), and South-West (SW, yellow). Shades light to dark indicate frequencies between 0–0.1%, 0.1–1%, and 1–10%. Pie chart shows the overall frequency in each quadrant and the total number of point reports (108903). Panel (b) shows the annual timeseries (1950–2024) of point flash flood reports: all flood types (grey bars), flash floods (dark red), and flash floods with latitude (lat)/longitude (lon) coordinates (red). Panel (c) shows the 2021 timeseries of point (red) and gridded (black) flash flood reports (over 24-hourly accumulation periods ending at 00 UTC). The blue circle highlights the reports within the period ending 2021-09-02 00 UTC (Storm Ida). Panel (d) show the spatial distribution of point (red dots) and gridded (black) impact reports for that same 24-hourly period. Zoomed area shows reports around New York City.

recordings will include impact severity, affected areas, and casualty information. Among all available keys[5], this study considers only those described in Table 4.1. It is worth noting that, as with any other human augmented reporting system, the Storm Event Database is subject to variations in other variables. Marjerison et al. (2016) highlight that population density is one of the most important factors affecting the location of flash flood reports. Figure 4.2a shows indeed that the Eastern side of the CONUS - the most densely populated - presents a larger frequency of reports than the Western side. Other factors affecting the location of flash flood reports are diurnal cycles of human activity and more mundane transcription or memory errors that affect the timing and location of reports (Barthold et al., 2015). Evidence of these issues has been found in assessments of Flash Flood Guidance (FFG) skill (Clark et al., 2014).

Storm Event Database: relevance for flash flood prediction and evaluation

Notwithstanding the acknowledged limitations in the flash flood event reporting in the Storm Event Database, this dataset demonstrates substantial utility for model development and evaluation compared to global datasets. The data preserves fundamental hydro-climatological signals essential for model training. Figure 4.2c (red bars), representative of flash flood report counts for 2021, shows that the underreporting does not compromise the representation of seasonal variability throughout the year - the timeseries shows pronounced peaks during the summer months and the beginning of the autumn when peaks in (flash) flood frequency are expected (Dougherty and Rasmussen, 2019). Furthermore, the database maintains sufficient daily report counts over most of the year. This ensures the dataset provides adequate sample sizes for model training and to compute robust statistics during model evaluation.

Storm Event Database: point-to-gridded post-processing of flash flood reports and their impact on model training and evaluation

As anticipated in Chapter 3, the point flash flood impact reports are post-processed into gridded fields so that they can be compared with the gridded reanalysis and forecast data. Whilst further details about the post-processing approach are provided in Section 5.2 in Chapter 5, this section focuses on the following observation. Days exhibiting exceptionally high report counts, such as those associated with Storm Ida (red bar within blue circle in Figure 4.2c), typically reflect spatially concentrated events rather

[5]Please, refer to https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Bulk-csv-Format.pdf for a detailed description of the features (columns) included in the Storm Event Database

Table 4.1: Features used from the Storm Event Database. The table presents the name of the features, called ”Keys” in the database, the variable types (e.g., string, data object, or float), their units, and description.

Key	Type	Units	Description
EVENT_ID	String	-	ID assigned by NWS for each individual storm event contained within a storm episode.
STATE	String	-	The (spelt out) name of the state where the event occurred.
CZ_TIMEZONE	String	-	Time Zone for the County/Parish, Zone or Marine Name. Eastern Standard Time (EST), Central Standard Time (CST), Mountain Standard Time (MST), etc.
SOURCE	String	-	The source reporting the weather event (e.g., Public, Newspaper, Law Enforcement, Broadcast Media, ASOS, Park and Forest Service, Trained Spotter, CoCoRaHS, etc.) . It can be any entry as it is not restricted to what is allowed.
EVENT_TYPE	String	-	Type of events (e.g., riverine or flash floods). The only event types permitted in SED are listed in Table 1 of Section 2.1.1 of the NWS Directive 10-1605.
FLOOD_CAUSE	String	-	Reported or estimated cause of the flood.
BEGIN_DATE_TIME	Date and Time (MM-DD-YYYY hh:mm:ss 24-hour format)	UTC	Date and time of the beginning of the flash flood event (e.g., Ice Jam, Heavy Rain, Heavy Rain/Snow Melt).
END_DATE_TIME	Date and Time (MM-DD-YYYY hh:mm:ss 24-hour format)	UTC	Date and time of the end of the flash flood event.
BEGIN_LAT	Float	Decimal degrees	The latitude of the begin point for the event or damage path.
BEGIN_LON	Float	Decimal degrees	The longitude of the begin point for the event or damage path.
END_LAT	Float	Decimal degrees	The latitude of the end point for the event or damage path.
END_LON	Float	Decimal degrees	The longitude of the end point for the event or damage path.

than widespread flash flood occurrences across the CONUS (Figure 4.2d). The point-to-grid post-processing is designed to prioritise the accurate representation of flash flood occurrence patterns due to hydro-meteorological factors. As a consequence, the approach treats all grid-boxes containing flash flood reports equally - a grid-box is assigned the ”yes-event” label regardless of whether it contains one or multiple point flash flood reports. The result of this methodological approach is exemplified by the height of the black bar within the blue circle in Figure 4.2c being much lower than the red bar. Whilst high report densities contain valuable information for modelling reporting behaviour, incorporating this information into this study could lead the model to learn unwanted patterns driven by factors external to hydro-meteorological processes. For example, report density variations may reflect the nature of the triggering event (hurricane vs. isolated convection, with the first type typically spurring people to contribute more reports in impact databases) or demographic factors (events in densely populated areas naturally produce more reports).

4.3 ERA5

ERA5 reanalysis and forecasts: general description

ERA5 represents the fifth-generation atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides hourly estimates of numerous atmospheric, land, and oceanic variables from 1940 to present, at approximately 31 km horizontal resolution. This comprehensive reanalysis combines vast quantities of historical observations with advanced numerical weather prediction models through sophisticated data assimilation techniques, yielding physically consistent and spatially complete reconstructions of past weather and climate conditions globally (Hersbach et al., 2020). The ERA5 system also produces near-real-time forecasts up to day 10. Table 4.2 describes the characteristics of both datasets (e.g., spatial and temporal resolution, run times, and maximum lead time).

ERA5 variables used to compute data-driven model features to identify and predict areas at risk of flash floods

ERA5 reanalysis was used to compute the hydrological and static features used to train the data-driven models developed in Chapters 6 and 7 to *identify* areas at risk of flash floods. During the inference stage, ERA5 forecasts up to day 5 served as inputs to the trained data-driven models to *predict* areas at risk of flash floods. From the extensive suite of variables

Table 4.2: ERA5 reanalysis and forecasts. The table describes the characteristics of the ERA5 reanalysis and forecasts.

Dataset	Runs (UTC)	Max Lead Time (hours)	Temporal Resolution	Version in Mars Archive
Reanalysis*	06 and 18	t + 18	Hourly	1
Forecasts*	00 and 12+	t + 240	3-hourly up to t+12 6-hourly up to t+120 12-hourly up to t+240	11

* **Model Version:** IFS Cycle 41r2, **Native spatial resolution:** reduced-gaussian N320 (~31 km at the equator), **Temporal range:** 1940 to near real time (5-day latency), **Mars Catalogue:** https://apps.ecmwf.int/mars-catalogue/?class=ea&stream=oper
+ The 12 UTC runs in the ERA5 forecasts were not used in this thesis.

available in ERA5[6], the subset of dynamic and static parameters listed in Table 4.3 were considered to compute antecedent soil moisture, orography slope, and vegetation coverage.

The following sub-sections describe the pre-processing of the raw ERA5 parameters to compute the variables that will be included in the data-driven model.

4.3.1 Parameter representing the antecedent soil moisture: percentage of soil maximum saturation

Percentage of soil maximum saturation (representing antecedent soil moisture): general description

To estimate the antecedent soil moisture, this study considers the percentage of soil maximum saturation in the top 1 metre layer of soil, 24 hours prior to a flash-flood-triggering rainfall event. This 1 metre depth corresponds to the active ”root zone” represented by the first three layers of the ECMWF’s IFS land surface scheme, defined as Layer 1, between 0 and 7 cm, Layer 2, between 7 and 28 cm, and Layer 3, between 28 to 100 cm (ECMWF, 2016). Table 4.4 shows the pre-defined values of maximum soil saturation used at ECMWF for different types of soil (Balsamo et al., 2009):

Equations 4.1 and 4.2 show how the fields containing the percentage of soil saturation were computed:

[6]The full set of parameters available for ERA5 reanalysis and forecasts can be found in the Mars Catalogue at `https://apps.ecmwf.int/mars-catalogue/?class=ea&stream=oper`.

Table 4.3: Parameters used from ERA5 and ERA5-ecPoint. Description of the parameters used to compute the hydrological and static features used in the developed data-driven models to identify and predict areas at risk of flash floods.

Name Parameter	Symbol	Range of values	Units	Mars ID	Type	Accumulation
Volumetric soil water, layer 1 (0 – 7 cm)	swvl1	Float, > 0	$m^3\ m^{-3}$	39	Dynamic	Instantaneous
Volumetric soil water, layer 2 (7 – 28 cm)	swvl2	Float, > 0	$m^3\ m^{-3}$	40	Dynamic	Instantaneous
Volumetric soil water, layer 3 (28 – 100 cm)	swvl3	Float, > 0	$m^3\ m^{-3}$	41	Dynamic	Instantaneous
Soil type	slt	Integer, from 0 to 7*	-	43	Static	n/a
Standard deviation of filtered sub-grid orography	sdfor	Float, > 0	m	74	Static	n/a
Slope of sub-grid orography	slor	Float, 0 to 1 (for 0 to 90 degree slope)	-	163	Static	n/a
Leaf area index, low vegetation	lai_lv	Float, 0 (bare soil) to 7 (dense canopy)	$m^2\ m^{-2}$	66	Dynamic, with no inter-annual variability	Instantaneous
Leaf area index, high vegetation	lai_hv	Float, from 0 (bare soil) to 7 (dense canopy)	$m^2\ m^{-2}$	67	Dynamic, with no inter-annual variability	Instantaneous
Low vegetation cover	cvl	Float, 0 to 1	-	27	Static	n/a
High vegetation cover	cvh	Float, 0 to 1	-	28	Static	n/a

* The **soil type codes** are coarse (code = 1), medium (2), medium fine (3), fine (4), very fine (5), organic (6), and tropical organic (7).

Table 4.4: Maximum soil saturation The table shows the pre-defined values of maximum soil saturation per soil as suggested by Balsamo et al. (2009).

Soil Type Code	1 Coarse	2 Medium	3 Medium Fine	4 Fine	5 Very fine	6 Organic	7 Tropical Organic
Maximum Saturation (-)	0.403	0.439	0.430	0.520	0.614	0.766	0.472

$$\text{max\_sat\_field} = \sum_{i=01}^{N} \text{max\_sat}_i \, \mathbf{1}_{\{s=c_i\}} \tag{4.1}$$

$$\mathbf{1}_{\{s=c_i\}} = \begin{cases} 1, & s = c_i, \\ 0, & s \neq c_i. \end{cases} \tag{4.2}$$

where, N goes from 1 to 7, and it represents the soil type codes. Each grid box is assigned a single, dominant soil type. The indicator function $\mathbf{1}_{\{s=c_i\}}$ acts as a binary filter: it returns 1 only when the grid box's soil type s matches a specific category c_i; it returns 0 otherwise.

Equation 4.3 computes the soil water content (swvl) over the top 1 metre layer of the soil integrating the soil water content over the top three layers:

$$\text{swvl} = \sum_{j=1}^{M} \text{swvl}_j \, \text{depth}_j \tag{4.3}$$

where M goes from soil layer 1 to 3. The percentage of the soil maximum saturation is considered a dynamic field because the values of swvl in the three soil layers change at every reanalysis and forecasts run[7].

Finally, the equation 4.4 computes the percentage of soil saturation as follows:

$$\text{swvl\_perc} = \frac{\text{swvl}}{\text{max\_sat\_field}} \tag{4.4}$$

[7]To know how to retrieve the fields for swvl, please, refer to the Mars Catalogue at `https://apps.ecmwf.int/mars-catalogue/?class=ea&stream=oper` .

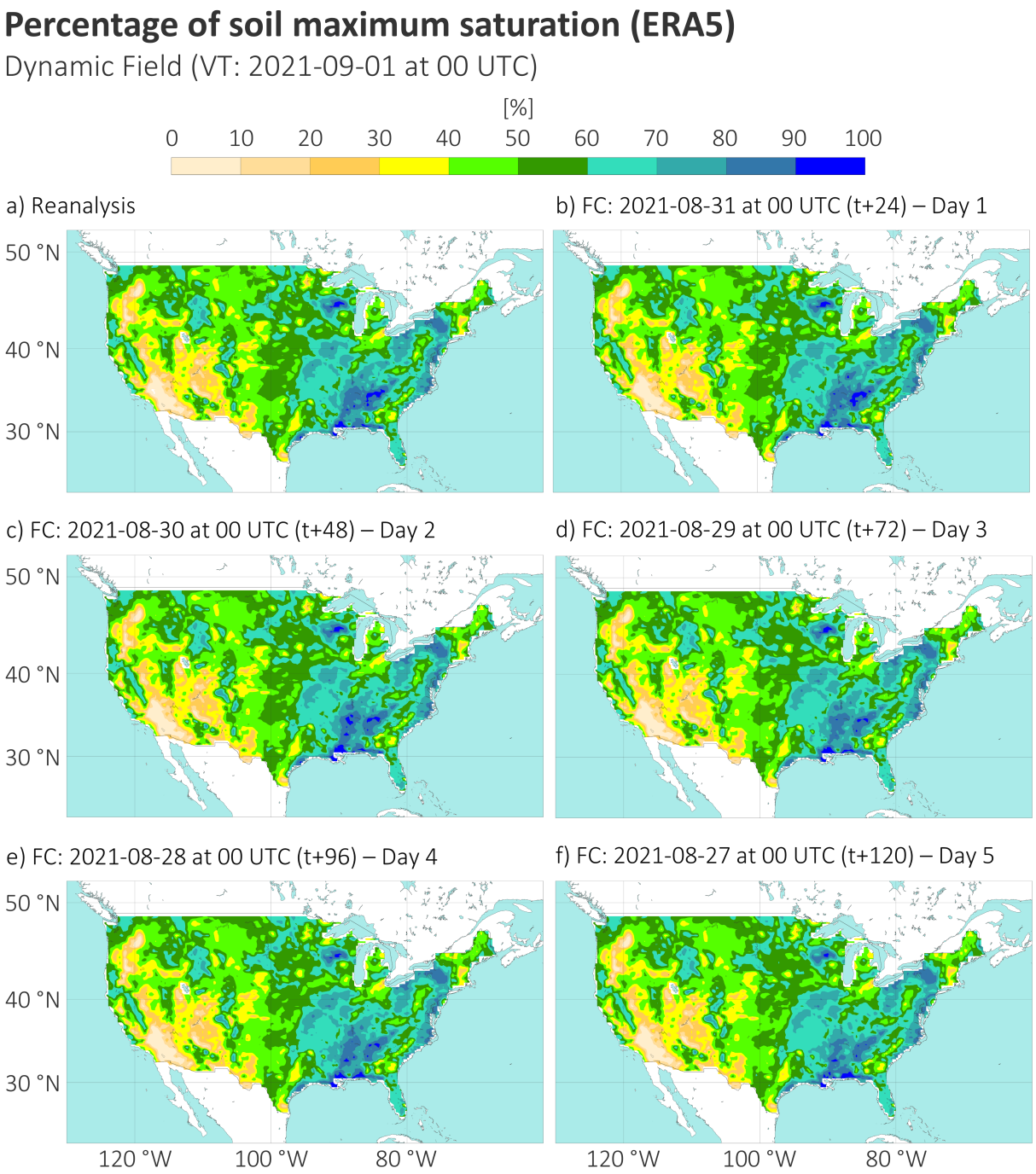

Figure 4.3: Percentage (%) of maximum soil saturation for the valid time (VT) corresponding to 2021-09-01 at 00 UTC Panel (a) shows the percentage of soil saturation in ERA5 reanalysis computed on 2021-08-31 at 18 UTC (t+6). Panels (b) to (f) show ERA5's forecasts computed, respectively, on 2021-08-31 at 00 UTC (t+24) - day 1, 2021-08-30 at 00 UTC (t+48) - day 2, 2021-08-29 at 00 UTC (t+72) - day 3, 2021-08-28 at 00 UTC (t+96) - day 4, and 2021-08-27 at 00 UTC (t+120) - day 5.

Example of the percentage of soil maximum saturation after Storm Ida

Figure 4.3 illustrates the spatial distribution of the percentage of soil maximum saturation across the CONUS for the valid time of September 1, 2021, at 00 UTC. Panel (a) presents the reference state derived from ERA5 reanalysis. Panels (b) to (f) display ERA5 forecasts for the same valid time but for different lead times, from t+24 (day 1) to t+120 (day 5). The colour scale indicates soil wetness, with warm colours (brown/yellow) representing dry conditions and cool colours (green/blue) indicating higher saturation. Heavy rainfall from Storm Ida, over the preceding days, caused a distinct area of near-total saturation (approaching 100%) in the southeastern CONUS. The figure demonstrates the model's ability to capture this saturation pattern up to five days in advance, as the forecast fields remain visually consistent with the reanalysis.

4.3.2 Parameter representing the orographic steepness: the standard deviation of the filtered sub-grid orography

Standard deviation of the filtered sub-grid orography (representing orographic steepness): general description

This study considers the standard deviation of the filtered sub-grid orography as the parameter to represent the orographic steepness. This is a static, time-invariant field, representing the statistical variability of terrain elevation at higher spatial resolution (typically at 1 km) than the model grid resolution (for ERA5, $\sim$31 km). It serves as a fundamental component for parametrising unresolved topographic effects in both atmospheric and hydrological modelling. The parameter is expressed in metres. In regions with high values of this parameter (typically above 100-500 meters), the parameter indicates significant topographic heterogeneity within model grid-boxes, representing the presence of valleys, ridges, peaks, and other terrain features that cannot be explicitly resolved. Figure 4.4 shows the standard deviation of the filtered sub-grid orography values over the CONUS.

Correlation and distinction between the orography's absolute elevation and the standard deviation of the filtered sub-grid orography.

Figure 4.1 and 4.4 are visually similar because high-elevation regions, such as the Rocky Mountains on the West and the Appalachians on the east, are geologically correlated with complex, rugged terrain. However, they indicate distinct physical properties: the orography defines the absolute vertical position relative to sea level, whereas the standard deviation of the filtered sub-grid orography quantifies the spread of elevation values around that mean, effectively representing terrain roughness. For instance, a high-altitude plateau would have high orography but low standard deviation because it is flat.

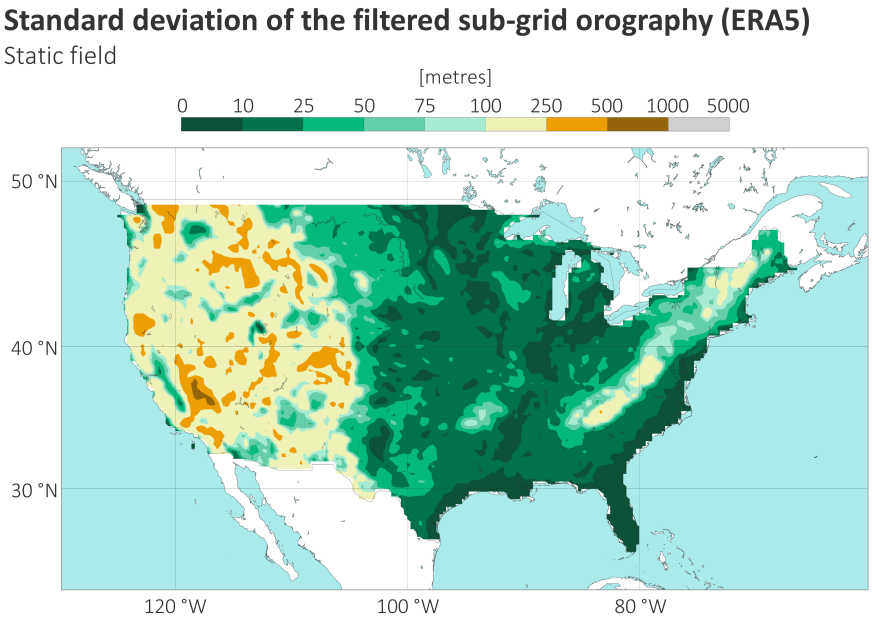

Figure 4.4: Standard deviation of the filtered sub-grid orography (static field). The figure presents the map plot showing the values for the standard deviation of the filtered sub-grid orography in ERA5 over the CONUS.

Justification on the selection of the standard deviation of the filtered sub-grid orography to represent orographic steepness over orography's slope

The standard deviation of sub-grid orography is preferred over grid-scale slope because it preserves the signal of local topographic roughness that controls rapid runoff, which is otherwise lost when calculating an average slope at the coarse model resolution (∼31 km). While a simple slope calculation would smooth out narrow valleys and ridges, the standard deviation metric explicitly quantifies this unresolved steepness.

4.3.3 Parameter representing the vegetation coverage: the leaf area index (LAI)

Leaf area index (representing vegetation coverage): general description

The leaf area index corresponds to a non-dimensional number representing the square metres of leaf area per square metre of the earth's surface[8]. It has a value of 0 over bare ground or where there are no leaves, and it grows as the vegetation coverage increases, typically up to values equal to 7. In ERA5, as in the ECMWF IFS, the leaf area index varies only climatologically, month by month. Hence, anomalous weather (e.g. winds

[8]https://confluence.ecmwf.int/display/FUG/Section+2.1.4.7+Modelling+vegetation.+Leaf+area+index

stripping leaves from trees or widespread fire damage) has no effect on its value. The leaf area index is calculated using the following formula:

$$LAI_{total} = (LAI_{high} \times Cover_{high}) + (LAI_{low} \times Cover_{low})$$

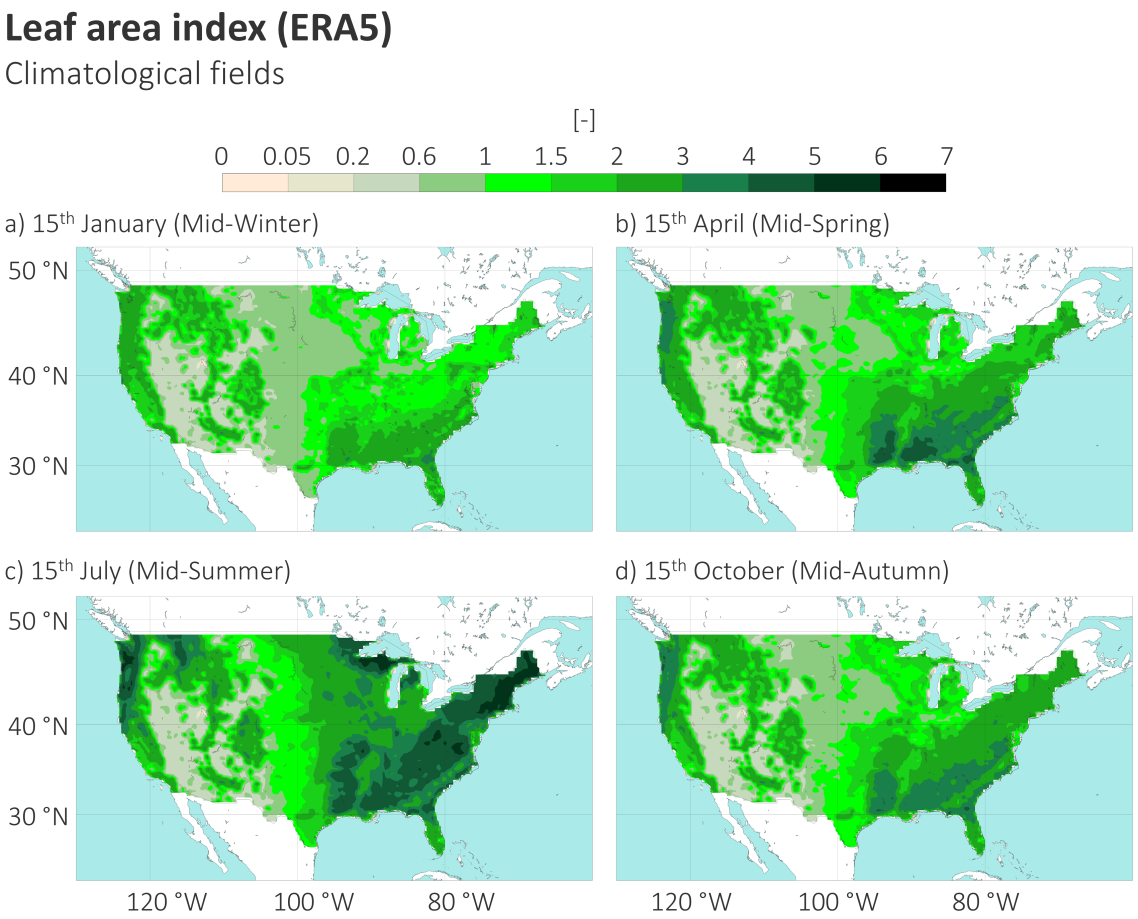

Figure 4.5: Leaf area index (climatological field). Panels (a) to (d) show examples of leaf area index values in ERA5 over the CONUS, respectively, for a day in mid-winter (15th of January), mid-spring (15th of April), mid-summer (15th of July), and mid-autumn (15th of October).

Seasonal evolution of the LAI values, reflecting the climatological phenological cycle of the vegetation across the CONUS

An example of LAI values over the CONUS during the four seasons is shown in Figure 4.5. In mid-winter (15th of January), LAI is at its minimum. Most interior and northern regions show values near 0–0.2, indicating dormant vegetation or bare ground, while evergreen forests in the Pacific North-West retain moderate coverage. By mid-spring (15th of April), a green-up is evident in the South-East where values rise to ranges between 2 and 4. The cycle peaks in mid-summer (15th of July), where vast areas of the eastern US and Pacific North-West reach saturation values ranging between 5 and

7. Finally, mid-autumn (15th of October) marks the decline associated with leaf senescence, as values in northern latitudes range only between 1 and 2, although the South-East retains relatively high vegetation coverage.

4.4 ERA5-ecPoint: post-processed ERA5 rainfall with the ecPoint post-processing technique

Limitations of ERA5's raw rainfall estimates (reanalysis or forecasts) in the prediction of flash-flood-triggering rainfall events

ERA5's raw rainfall estimates (reanalysis or forecasts) exhibit known limitations in representing rainfall extremes, which are detrimental for the correct prediction of flash flood events. As any other gridded model output, ERA5 provides a rainfall estimate that represents the average of point values over the grid-box area. This averaging inevitably smooths the localised, high-intensity rainfall peaks, which are the primary drivers of flash floods.

Description of the ecPoint post-processing technique

ecPoint is a statistical post-processing technique that transforms global gridded NWP outputs into probabilistic point-scale forecasts (Hewson and Pillosu, 2021). The post-processing technique aims to provide post-processed forecasts that mirror observations from rain gauges by addressing the two main factors affecting the performance of global NWP model outputs against point verification: systematic biases (Lavers et al., 2021) and lack of representation of sub-grid variability (Göber et al., 2008). The errors between global gridded rainfall forecasts (i.e., up to t+30, control run of ECMWF's ENS) and point-rainfall observations (i.e., rain gauges) are computed for a one-year calibration period. The error computed for accumulated variables (like rainfall) is the Forecast Error Ratio (FER), whose formulation is shown in Figure 4.6a. The error distribution is named Mapping Function (MF), and its shape is linked to the degree of sub-grid variability and biases at grid scale in the raw forecasts. The MF for all data points in the calibration period is also shown in Figure 4.6a, and it shows that ECMWF's ENS both overestimates (green bars) and underestimates (yellow and red bars) versus gauge reports. The white bar indicates that only ~15% of the point-rainfall observations were correctly predicted.

Difference between a univariate and multivariate approach to post-process rainfall outputs from a global NWP model.

The MF is used to post-process the raw rainfall outputs from NWP models. Suppose all grid-boxes in the raw forecasts are post-processed by sampling only the MF in Figure 4.6a (also the same MF in the black circle in Figure 4.6b). In this case, the ecPoint post-processing would follow a univariate approach (U-ecPoint), and be thought of as a single-leaf

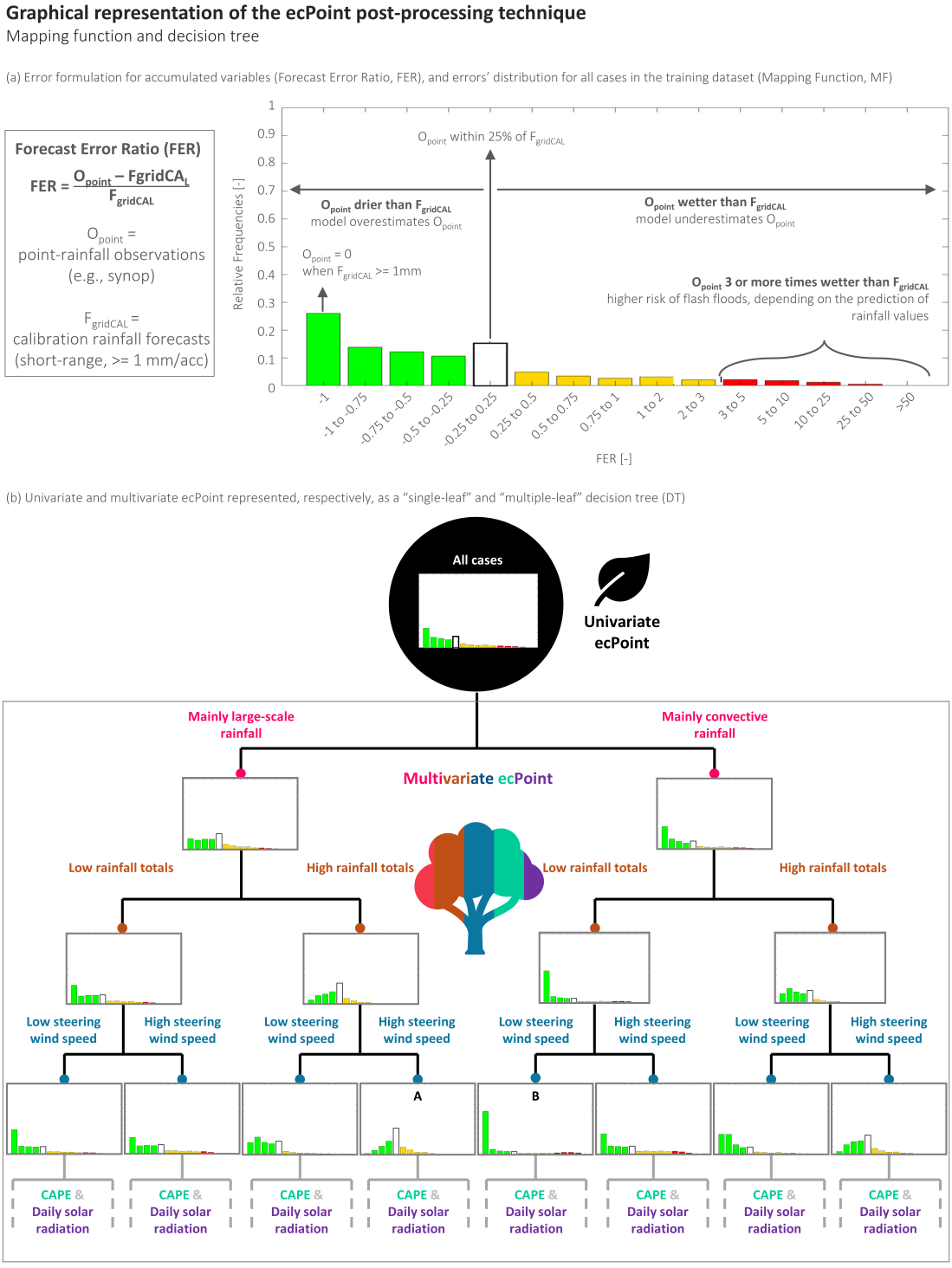

Figure 4.6: Graphical representation of the ecPoint post-processing technique, from (Pillosu et al., 2025a). Panel (a) shows the error formulation for accumulated variables (Forecast Error Ratio, FER) and the error distribution for all cases in the training dataset (Mapping Function, MF). The example pertains to the calibration of 47r3 ECMWF ENS forecasts for 12-hourly rainfall forecasts. Panel (b) shows the univariate approach for ecPoint (U-ecPoint) represented as a "single-leaf" decision tree (DT, within the black circle), while the multivariate approach (M-ecPoint) is represented as a "multiple-leaf" DT (within the grey square).

decision tree (Figure 4.6b). U-ecPoint generally increases the number of zeros in the distribution of point-rainfall forecasts to correct ENS's tendency to overpredict small rainfall totals (Haiden et al., 2023). It also increases the rainfall amounts in the wet tail of the rainfall distribution to correct for ENS underestimation of high rainfall values (Haiden et al., 2023). The MF shape can, however, change significantly according to different weather scenarios at grid-scale (called Grid-box Weather Type, G-WT). The multiple MFs can be visualised with a decision-tree-like representation, where each leaf of the decision tree corresponds to a G-WT and its corresponding MF (Figure 4.6b. When each grid-box in the raw forecast is post-processed differently by using the MF corresponding to the matching G-WT in the decision tree (within the grey square in Figure 4.6b), the ecPoint post-processing follows a multivariate approach (M-ecPoint). It corresponds to the original ecPoint system developed by Hewson and Pillosu (2021). When for a grid-box, the raw ENS predicts high totals of mainly large-scale rainfall and strong steering wind speeds (case A in Figure 4.6b), the MF takes a Gaussian-like form. This means the raw model output is relatively representative of the point-scale rainfall totals. When the raw ENS predicts mainly convective rainfall with light steering wind speeds (case B in Figure 4.6b), the MF might take an exponential-like form. This means that the raw model output is not representative of the point-scale rainfall totals and that the expected degree of sub-grid variability is bigger than in case A. Each raw forecast is converted into a distribution of N point-scale forecasts using the MFs (for example, operationally, for each raw ensemble member, N=100 point-scale forecasts are created). Hence, while M-ecPoint increases overall the frequency of small and large rainfall totals in the post-processed forecasts, as U-ecPoint does, its adjustments are applied according to different G-WTs. M-ecPoint reduces the probabilities at certain locations more than U-ecPoint; this relates to the fact that corrections are applied differently across the ensemble members rather than uniformly, as done by U-ecPoint. Moreover, Hewson and Pillosu (2021) have shown that, due to the G-WT differentiation in the corrections, one of M-ecPoint's features is the ability to shift the location of areas at higher risk of extreme localised rainfall. This feature is lost in U-ecPoint as all grid-boxes are post-processed identically (Pillosu et al., 2025a).

ERA5-ecPoint

The ecPoint post-processing technique has been applied to ERA5 reanalysis (Bottazzi et al., 2024). The deterministic realisations of ERA5 reanalysis and forecasts are, therefore, transformed to a distribution of 100 point-rainfall totals, and distilled in 99 percentiles from 1st to 99th. ecPoint

reanalysis and forecasts are provided in the same native grid of ERA5 (reduced Gaussian grid N320, $\sim$31 km), with forecasts up to day 5, and accumulation periods ending at 00 UTC. Preliminary verification between raw ERA5 and ERA5-ecPoint reanalysis has shown that ERA5-ecPoint represents point rainfall estimates better than ERA5 (Hewson et al., 2023). Subsequently, Pillosu et al. (2025b) carried out a more systematic verification analysis against point-scale rainfall climatologies from rain gauges scattered worldwide. Compared to raw ERA5, Pillosu et al. (2025b) showed that ERA5-ecPoint rainfall estimates provide a better performance in the representation of the overall distribution of point-scale rainfall distribution, including extreme rainfall exceeding the 10-year return period.

Example for the probabilities of ERA5-ecPoint rainfall exceeding the 1-year and 50-year return period

Figure 4.7 and 4.8 show examples of the probabilities of 24-hourly rainfall exceeding, respectively, the 1-year and 50-year return period in ERA5-ecPoint reanalysis and forecasts. The north-east coast of the CONUS was affected by Storm Ida, a large-scale convective system. For this event, the probabilities of exceeding both the 1-year and 50-year return periods are uniformly elevated across an extensive area, suggesting a spatially widespread event. The signal of widespread extreme rainfall in the area remains robust (with probabilities exceeding 60%) all lead times, demonstrating strong predictability for this kind of system. The south-west quadrant of the CONUS was affected by heavy rainfall from a storm-scale convective system. At day 1, the probabilities of exceeding the 1-year return period are on average 20% while, at day 5, they reduce to 5%. Moreover, the location of the localised extremes exhibits considerable spatial variability across successive runs. A similar behaviour is observed for the rainfall forecasts exceeding the 50-year return period. In this latter case, predicting the correct areas that may experience intense rainfall is more difficult, as the locations vary significantly over time.

4.5 Case Study used throughout the thesis: Storm Ida

Storm Ida: synoptic history

Tropical Storm Ida formed on 23 August 2021 in the western Caribbean Sea, southwest of Jamaica[9]. It first hit western Cuba on 27 August as

[9]National Hurricane Center best track of Hurricane Ida (https://www.weather.gov/lch/2021Ida)

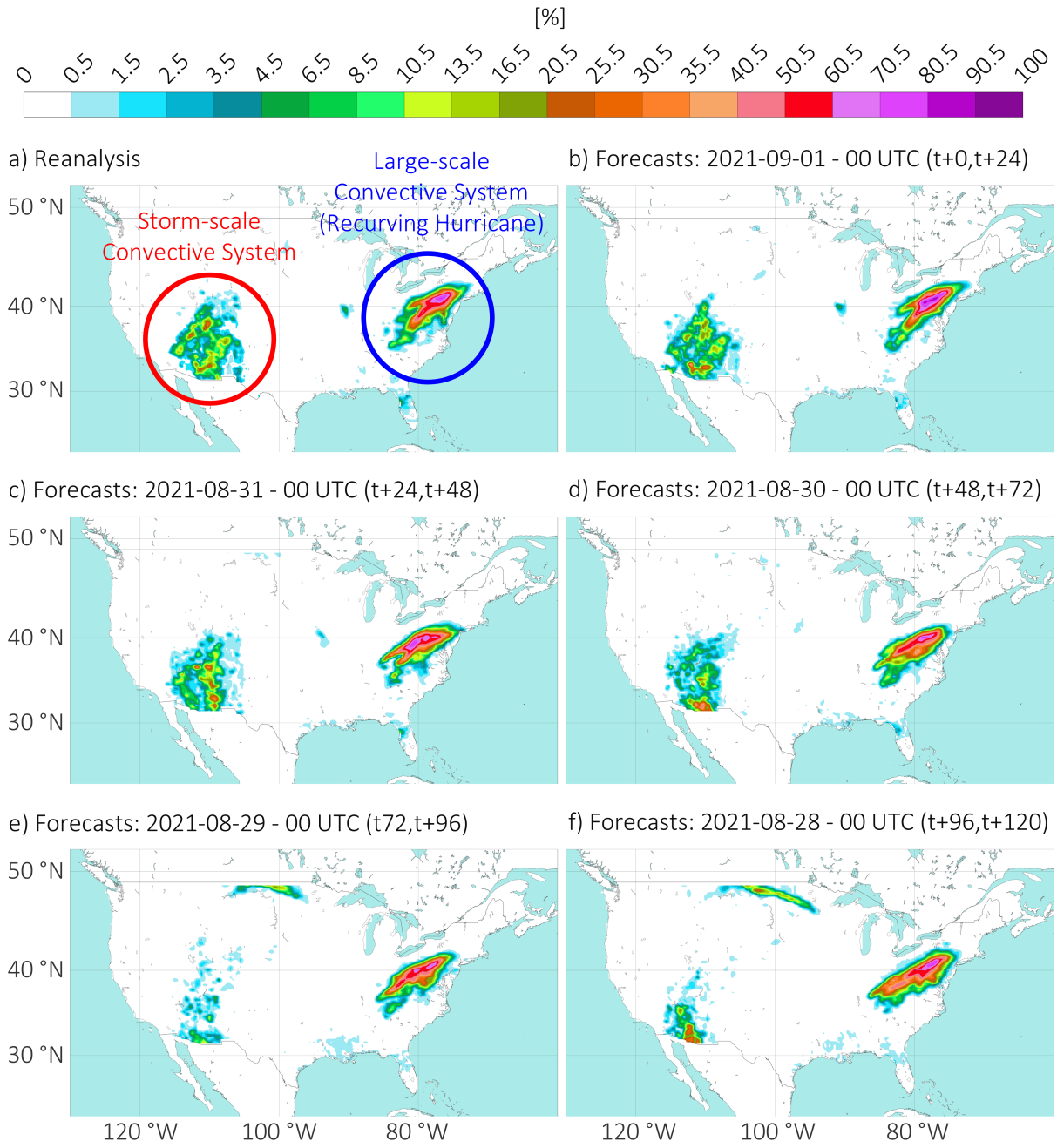

Figure 4.7: Probability (%) of exceeding the 1-year return period in ERA5-ecPoint. Panel (a) displays probabilities in ERA5-ecPoint reanalysis for the valid time (VT) ending on 2021-09-02 at 00 UTC. Panels (b) to (f) represent the probabilities in ERA5-ecPoint forecasts for the same VT, but for forecasts at day 1 (t+0,t+24), day 2 (t+24,t+48), day 3 (t+48,t+72), day 4 (t+72,t+96), and day 5 (t+96,t+120), respectively.

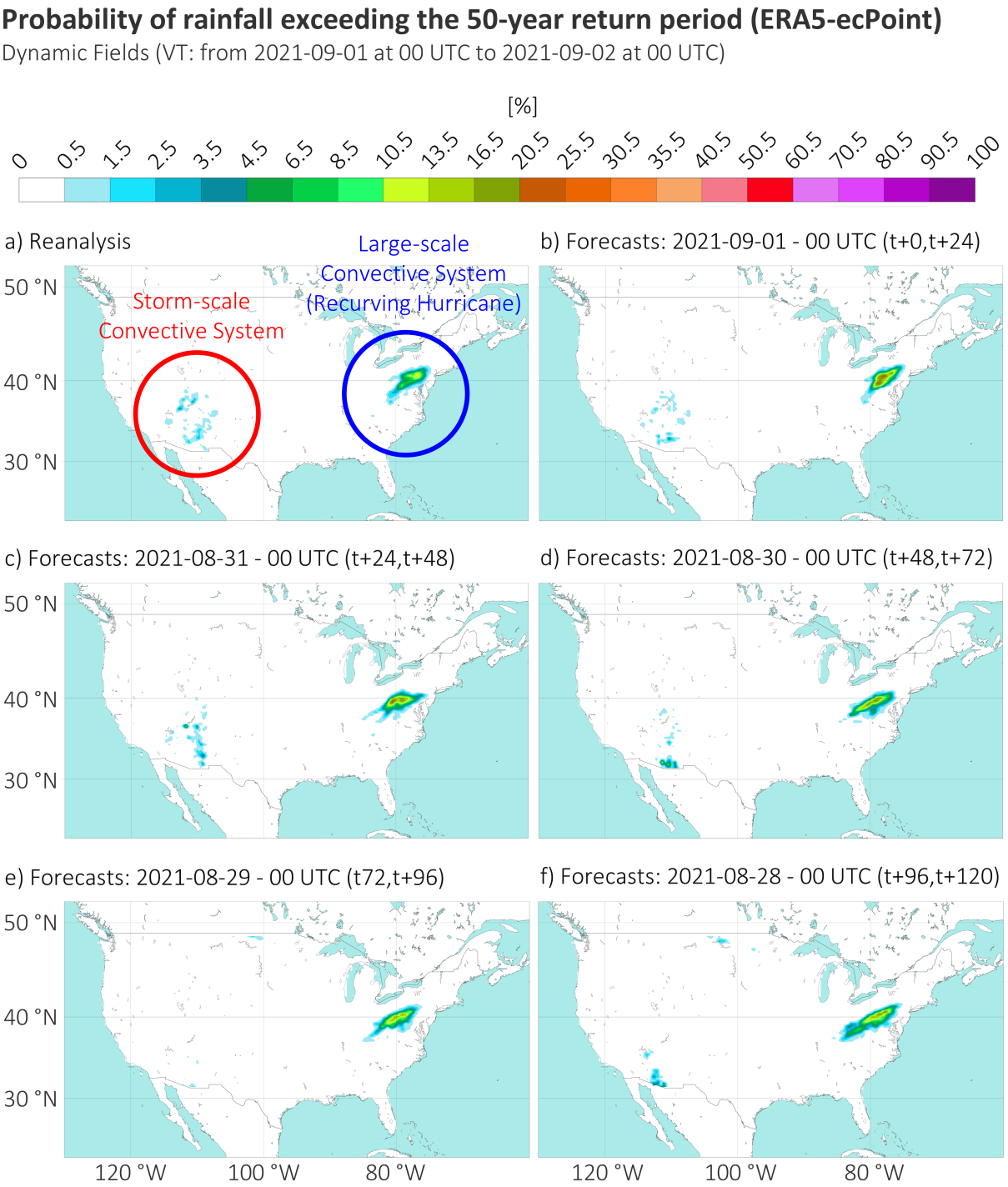

Figure 4.8: Probability (%) of exceeding the 50-year return period in ERA5-ecPoint. Similar to Figure 4.7, but for probability of exceeding the 50-year return period.

a Category 1 Hurricane (Figure 4.9). It then travelled northwestward towards the Gulf of Mexico, where it intensified rapidly due to sea surface temperatures of nearly 30°C. Ida made landfall on the Louisiana coast near Port Fourchon on 29 August 2021 at 16:55 CDT (Central Daylight Time, UTC-05:00 - 16:55 UTC) as a Category 4 Hurricane, with maximum sustained winds of 150 mph. The cyclone's intensity steadily decreased as it moved inland, and it weakened to a tropical storm before the centre moved into southwestern Mississippi between 06:00 and 12:00 UTC on 30 August 2021. Ida then turned northeastward as it moved around the western end of the subtropical ridge, with the centre passing just west of Jackson, Mississippi, around 18:00 UTC. Soon thereafter, the cyclone weakened to a tropical depression as it moved into northeastern Mississippi. The system then accelerated northeastward across northwestern Alabama, central and eastern Tennessee, and portions of Kentucky and Virginia before reaching southern West Virginia near 1200 UTC 1 September. Ida began extratropical transition as it moved through the Tennessee Valley, and the system became an extratropical low as it moved over West Virginia later that day. Once it became extratropical, Ida moved east-northeastward in the mid-latitude westerly flow through West Virginia, northern Virginia, and central Maryland to southeastern Pennsylvania by 0000 UTC 2 September. At that time, the system acquired gale-force winds over the Atlantic east of the centre. A continued east-northeastward motion brought the centre across northern New Jersey and into the Atlantic just south of Long Island, New York, to near Nantucket, Massachusetts by 1200 UTC that day. The low then turned northeastward and strengthened a little, reaching western Nova Scotia late on 2 September and moving into the Gulf of St. Lawrence on 3 September. This was followed by a cyclonic loop over the Gulf of St. Lawrence on 3–4 September while the low maintained maximum winds of 40–45 kt. The low degenerated to a trough late on 4 September as a new mid-latitude low formed to the east (Beven et al., 2022). The track followed by the storm is shown in Figure 4.9a.

Storm Ida: rainfall and flooding

As a tropical cyclone, Ida produced widespread heavy rain (up to ∼380 mm/24h) along portions of the northern Gulf coast states northward and eastward into the Tennessee Valley. This rain produced widespread riverine flooding, especially along the Tangipahoa, Tchefuncte, Tickfaw, and Bogue Falaya Rivers in southeastern Louisiana and the Tchoutacabouffa, Biloxi, Wolf, and Jourdan Rivers in southeastern Mississippi. When Ida became an extratropical cyclone, a swath of heavy rains (up to ∼250 mm/24h) developed north of the centre and affected a long area extending from northern

Case study analysed in the thesis

Storm Ida

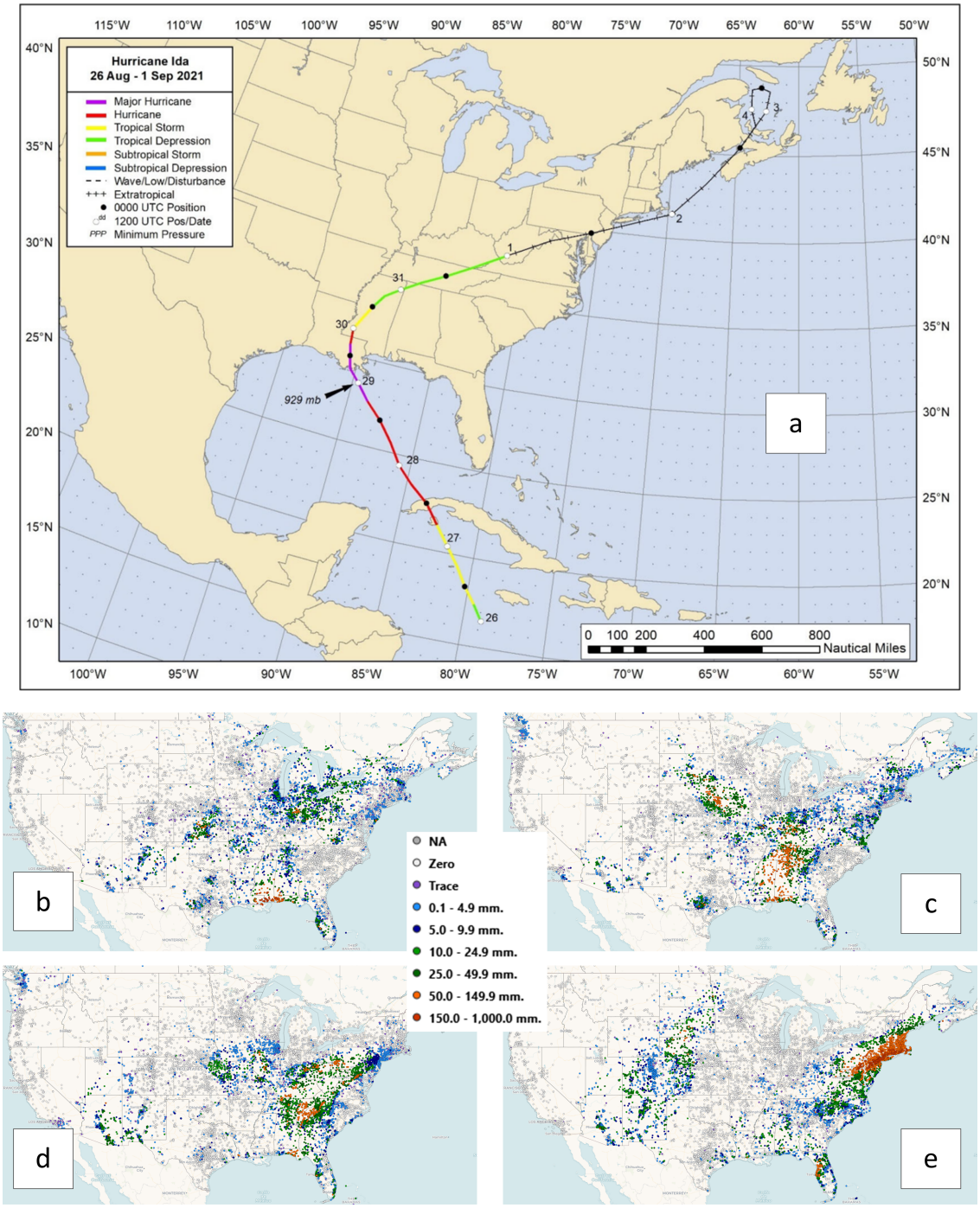

Figure 4.9: Case study analysed over the thesis - Storm Ida. Panel (a) shows the track followed by Storm Ida. Panels (b) to (e) shows 24-hourly rainfall totals over 2021-08-30, 2021-08-31, 2021-09-01, and 2021-09-02, between 4.30 am and 9.30 am local time. Observations were obtained from https://maps.cocorahs.org/

West Virginia, across western Maryland, southeastern Pennsylvania, northern New Jersey, southeastern New York, Connecticut, and Rhode Island to southeastern Massachusetts, including the New York City metropolitan area. The extreme rainfall rates and heavy rainfall caused major riverine flooding in these areas, including deadly and damaging flash flooding and urban flooding across portions of the New York City metropolitan area and northern New Jersey. Ida caused the fifth-wettest day in NYC history[10]. The observed rainfall totals during Storm Ida are shown in Figure 4.9b-e.

Storm Ida: impacts

Nearly all of southeast Louisiana lost power, leaving more than 1 million customers without electricity as the storm tracked northward through the state and into Mississippi. Even days later, on September 3, more than 900,000 remained without power. The storm devastated the Louisiana coast. Virtually every house on the barrier reef of Grand Isle was damaged, with many destroyed. Although the damage was immense, the good news was that the flood barriers in New Orleans held, so the city avoided the deadly flooding that had occurred during Hurricane Katrina in 2005. The band of torrential rains that stretched across New Jersey into the New York City area and Connecticut that Wednesday night triggered a rare flooding emergency. Flooding was so severe and sudden that basement apartment dwellers in New York City were overwhelmed by the floodwaters (LeComte, 2022). Ida was blamed for 96 deaths total and cost $75 billion, making it the 6th most costly storm to impact the US[11].

[10] https://www.ncei.noaa.gov/access/billions/dcmi.pdf

[11] https://www.ncei.noaa.gov/access/billions/dcmi.pdf

CHAPTER 5

FLASH-FLOOD-FOCUSED VERIFICATION OF RAINFALL-BASED PREDICTIONS OF AREAS AT RISK OF FLASH FLOODS

5.1 Introduction

On the importance of early warning systems to mitigate flash flood impacts

Flash floods[1] cause significant societal, economic, and environmental impacts (Dordevic et al., 2020). Forecast-triggered mitigation strategies, such as early warning systems (Coughlan De Perez et al., 2022; Šakić Trogrlić et al., 2022) and forecast-based financing protocols (Bischiniotis et al., 2019; Perez et al., 2016), have been shown to improve resilience, decrease mortality, and lower recovery costs against riverine floods. Yet, these strategies hinge on accurate, timely predictions. In lower-income countries, accurate forecasts with even longer lead times are required to set cost-effective mitigation strategies (Bazo et al., 2019; Kiptum et al., 2023).

Dependence on short-range, km-scale NWP models in regional, national, and continental flash flood forecasting systems, generating patchy spatial coverage and limited lead times)

Over the years, flash flood forecasting systems have been developed at local/regional (Speight et al., 2018; Corral et al., 2019; Ibarreche et al., 2020; Ramos Filho et al., 2021; Shuvo et al., 2021), national (Javelle et al., 2016; Liu et al., 2018), and continental scales (Gourley et al., 2017; Raynaud et al., 2015), with varying degrees of model complexity and forecast

[1]Please refer to the first paragraph in Chapter 1 for the definition of flash floods used in this thesis.

accuracy. These systems share a reliance on high-density rainfall and discharge observations, and km-scale rainfall forecasts (Braud et al., 2014). This approach has prevented the development of a flash flood forecasting system that provides medium-range forecasts over a continuous global domain. Radar-derived rainfall predictions remain limited to a few hours ahead (Imhoff et al., 2022), and kilometre-scale forecasts show substantial skill reduction beyond two days (Barrett et al., 2019). Moreover, the uneven spatial distribution of these data sources restricts flash flood prediction to a collection of separate regional and national systems, including WMO's "Flash Flood Guidance System with Global Coverage" (Georgakakos et al., 2022). Consequently, many areas of the world remain without access to flash flood guidance, highlighting the need for alternative approaches to address the challenge of providing medium-range predictions of areas at risk of flash floods over a continuous global domain.

Statistically post-processed rainfall forecasts from global NWP can provide suitable predictions for flash-flood-inducing rainfall events

Global NWP models, such as ECMWF's IFS, provide daily global rainfall predictions up to medium-range leads but have historically struggled to accurately predict extreme localised rainfall events due to their coarse resolution and parametrisation schemes (Emerton et al., 2016; Wen et al., 2021). Owing to recent improvements in global NWP forecast accuracy (Haiden et al., 2023; Lavers et al., 2021), the interest in using them to provide flash flood guidance in data-scarce regions and extend predictions' lead times has recently increased (Bucherie et al., 2022b). Additionally, the development of state-of-the-art post-processing techniques can enhance the quality of raw forecasts and make them more suitable for flash flood prediction (Vannitsem et al., 2021). For example, the ecPoint statistical post-processing technique transforms global grid-based forecasts into probabilistic point-scale predictions, improving the reliability and discrimination ability of rainfall forecasts up to day 10, especially for extremes (Hewson and Pillosu, 2021).

Aims of this chapter: bridge the evaluation gap of statistically post-processed rainfall forecasts for the identification of areas at risk of flash flood over a continuous global domain and up to medium-range lead times

However, a significant gap remains in leveraging the statistically post-processed rainfall forecasts from global NWP models to create truly global predictions of areas at risk of flash floods up to medium-range lead times. This study aims to bridge this gap by evaluating the performance of ERA5-ecPoint rainfall forecasts in identifying areas at risk of flash floods (from now on, this type of forecast will be referred to as *rainfall-based predictions of areas at risk of flash floods*). The CONUS, with its extensive collection of flash flood impact reports within NOAA's Storm Event Database and its varying climate, serves as an ideal test bed for this research. The innovation

proposed by this research is twofold. It proposes the first systematic flash-flood-focused verification framework for predictions of areas at risk of flash floods, in recognition of the non-linear relationship between flash floods and triggering rainfall events. Secondly, this research provides a performance benchmark against which more sophisticated modelling approaches (for example, incorporating additional hydrological and topographical parameters) can be measured. Such comparative analysis is essential for determining whether the increased computational demands and data requirements of complex systems yield commensurate improvements in flash flood prediction accuracy, or whether simpler precipitation-based approaches provide sufficient utility for early warning applications.

Chapter outline

The remainder of this chapter is organised as follows. Section 5.2 presents the development of the flash-flood-focused objective verification framework, describing the pre-processing of observational and forecast data, the construction of contingency tables for probabilistic forecasts using non-standard observations, and the verification scores employed to assess forecasts' reliability and discrimination ability. Section 5.3 presents the verification results, while Section 5.4 illustrates the results for the case study on Storm Ida, demonstrating how the verification metrics translate into practical forecast interpretation. Section 5.5 discusses the implications of these findings for operational flash flood guidance, with particular attention to the challenges posed by observation quality and event underreporting. Finally, Section **??** summarises the principal conclusions and outlines recommendations for future verification and further developments in the prediction of areas at risk of flash flood.

5.2 Development of the flash-flood-focused objective verification framework for rainfall-based predictions of areas at risk of flash floods

5.2.1 Pre-processing of observational and forecast data

Rainfall-based predictions of areas at risk of flash floods computed with ERA5-ecPoint rainfall forecasts: determining yes- and non-events in the forecast

ERA5-ecPoint rainfall forecasts (mm/24h at point-scale) - already described in Section 4.4 in Chapter 4 - are used to create the rainfall-based predictions of areas at risk of flash floods. ERA5-ecPoint is preferred to raw ERA5 estimates because they are able to capture the signal of flash-

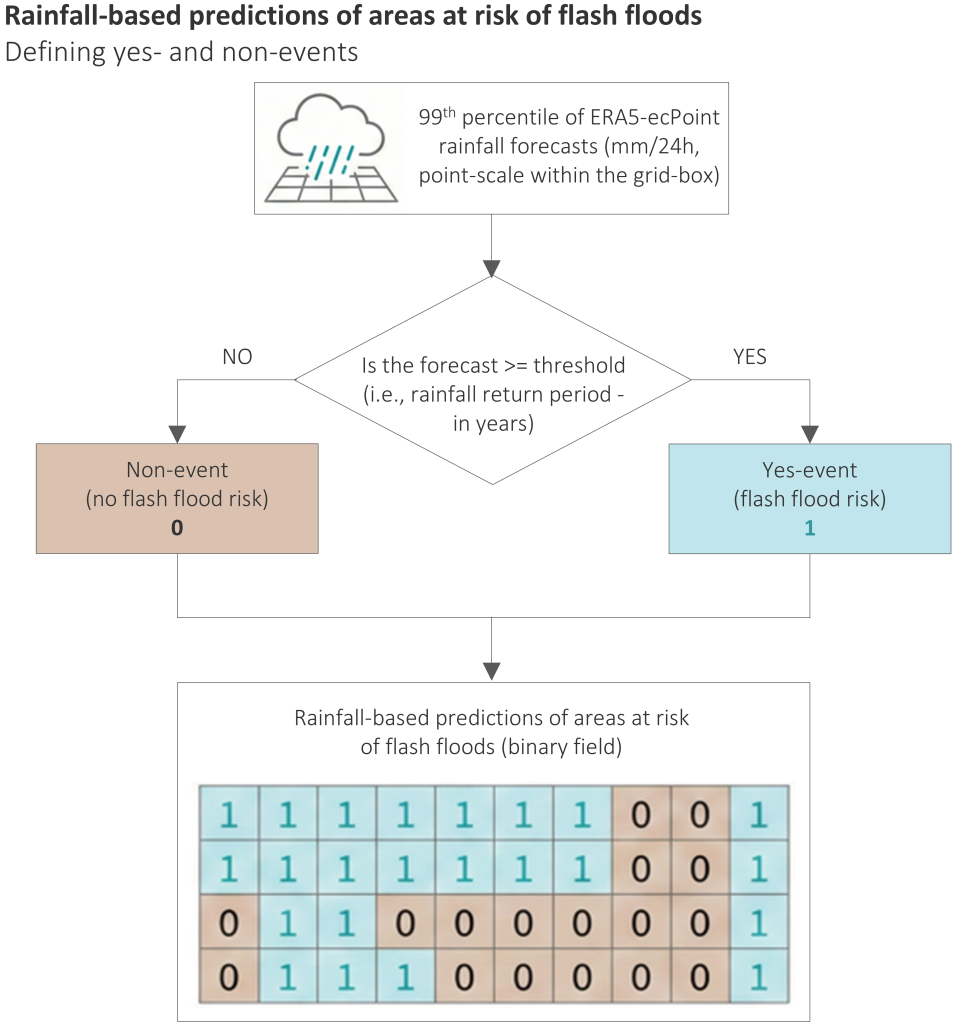

Figure 5.1: Schematic on how yes- and non-events are defined for rainfall-based flash flood predictions of areas at risk of flash floods. The point-scale forecast (mm/24h) for each grid box is compared against a specific rainfall threshold. If the forecast exceeds or is equal to the threshold, it is classified as a "yes-event" (value 1, shown in light green), indicating a risk of flash flooding. Conversely, forecasts below the threshold are classified as "non-events" (value 0, shown in light brown). The resulting output is a binary field representing the predictions of areas at risk.

flood-triggering rainfall. A grid-box is classified at a risk of flash flooding if the forecast exceeds a specific rainfall threshold (likely to generate a flash flood). In this case, the grid-box is assigned the value 1 (yes-event); 0 otherwise (non-event). The resulting binary field represents the rainfall-based prediction of areas at risk of flash floods (Figure 5.2).

Selecting the right dataset for the definition of the verifying rainfall thresholds

In objective verification, the threshold used to define yes- and non-events is called *verifying rainfall threshold*, and choosing the right dataset to compute it is critical. The verifying rainfall threshold must be defined in the same units and spatial resolution as the forecasts, so in our case, point-

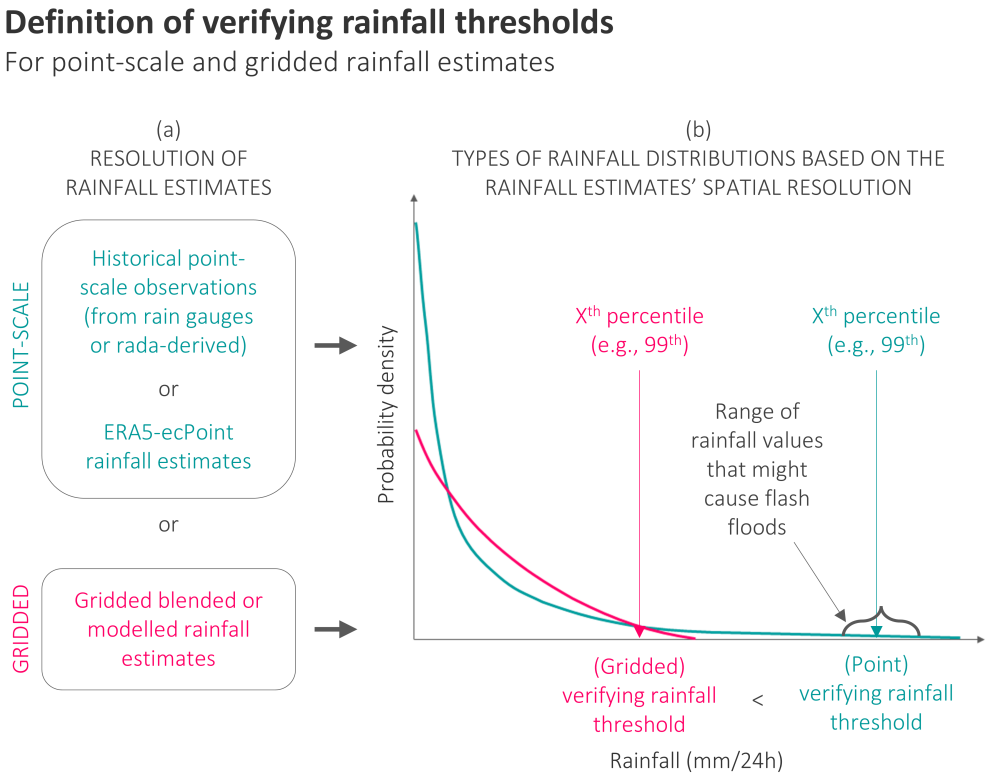

Figure 5.2: Schematic illustrating the process of defining a verifying rainfall threshold. Historical rainfall data from observations (in green) or gridded products (in pink) can be used to construct a probability distribution of rainfall intensity. An X^{th} percentile from the tail of this distribution, representing rare and intense events likely to cause flash flooding, is then selected to determine the verifying rainfall threshold. Verifying rainfall thresholds from a distribution constructed from gridded rainfall estimates is typically smaller than those computed using point-scale rainfall estimates.

scale rainfall expressed in mm/24h. Point-scale verifying rainfall thresholds of practical interest may be known (e.g., in my region, 50 mm/24h is likely to cause flash floods), or may be computed using historical observational or modelled point-scale rainfall estimates. Computing the verifying rainfall threshold from a long time series (at least 20–30 years) of high-density rain gauge rainfall observations (spaced only a few kilometres apart) or radar-derived rainfall totals (see Figure 5.2a, datasets in green) yields the most accurate representation of those localised rainfall extremes that are likely to trigger flash floods (Haiden and Duffy, 2016; Ramos Filho et al.,

2021)[2] (see schematic in Figure 5.2b, distribution in green). Due to the absence of a long timeseries of high-density rain gauge observations over the CONUS, the verifying rainfall thresholds were computed with point-scale rainfall estimates from ERA5-ecPoint reanalysis (Bottazzi et al., 2024), as they have been shown to represent the climatologies of rainfall observations from rain gauges (Pillosu et al., 2025b) adequately.

Computing the verifying rainfall thresholds: the mathematics

The verifying rainfall thresholds were computed over the 1991–2020 period, following WMO best-practice for climatological analysis (WMO, 2017). For each grid-box, a rainfall distribution was constructed from the 30-year timeseries. Verifying rainfall thresholds are defined as percentiles of this distribution: higher percentiles (typically, above the 99th percentile) correspond to rarer, more intense rainfall events that are more likely to produce flash flooding. Percentiles may also be expressed as return periods. For daily rainfall (mm/24h), the return period T (in years) is given by equation 5.1:

$$T = \frac{1}{365 \times (1-p)} \quad \text{[years]} \tag{5.1}$$

where, p is the percentile expressed as a probability (e.g. 0.99 for the 99th percentile). Table 5.1 shows the percentiles that are commonly used to define flash-flood-triggering rainfall events with different levels of severity and their equivalent in return periods (in years). These return periods will also be considered in this chapter.

Defining the gridded observational fields

To assess the performance of the rainfall-based predictions of areas

[2]In the absence of a suitable observational network, verifying rainfall thresholds could be defined from gridded rainfall products (see Figure 5.2a, datasets in pink), such as reanalysis, e.g., ERA5 (Hersbach et al., 2020), reforecasts (Hamill et al., 2006), or blended rainfall observations provided on a grid such as MSWEP (Beck et al., 2019) or GPCP (Adler et al., 2018). However, gridded rainfall data should not be used to compute verifying rainfall thresholds for flash flood applications. First, they tend to underestimate the ”true” value of flash-flood-triggering rainfall events, reducing users’ confidence in the methodology and its practical applicability (Tapiador et al., 2019). Moreover, as we are using point-scale rainfall forecasts to define the areas at risk of flash floods, the point-scale forecasts would exceed the verifying rainfall thresholds too often, giving the misleading impression that the forecasts overpredict the risk of flash floods, as there would be too many false alarms (see schematic in Figure 5.2b, distribution in pink).

Table 5.1: Severity of flash-flood-triggering rainfall events. The table shows the percentiles that define different severity categories for flash-flood-triggering rainfall events and their equivalent in return periods (in years, for mm/24h). Results for the starred (*) return periods are discussed but not shown in the "Results" Section 5.3.

Percentiles (in %)	99.726	99.945	99.973	99.986	99.995	99.997
Equivalent Return Period (in years)	1	5*	10*	20*	50	100*

at risk of flash floods, these must be evaluated against ground-truth observations. In this thesis, these ground truth observations are provided by point/polygon flash-flood impact reports from NOAA's Storm Events Database (refer to Table 4.1 in Section 4.2). The two datasets are not directly comparable because the forecasts are defined on a grid (reduced Gaussian grid N320 with approximately 31 km spatial resolution at the equator) and over an accumulated field (24-hourly, from 00 to 00 UTC), while the observations are provided at points at a specific time (instantaneous). Hence, the impact reports must be converted to the same grid and 24-hourly accumulation period of the forecasts. Practically, the impact reports are first grouped into the corresponding 24-hour accumulation periods (00–00 UTC). Each point/polygon report is then mapped to the model grid: in the case of a point report, this is assigned to the closest grid-box using the "nearest grid-box" method, while in the case of a polygon report, a report is assigned to all the grid-boxes within the polygon. The total number of reports accumulated in each grid-box is then counted. All the grid-boxes containing at least one flash flood report are assigned the value 1; 0, otherwise. Spatio-temporal buffers were not applied to the impact reports, unlike other studies using forecasts with finer spatial and temporal resolutions (Cavaiola et al., 2024). This choice is unlikely to affect the results as only 0.1% of all reports lie sufficiently close to a grid-box boundary or the 24-hour accumulation cut-off that omitting a buffer could plausibly shift the event into a neighbouring grid box or an adjacent accumulation period. We therefore make no additional adjustments for uncertainties in the reports' location or timing.

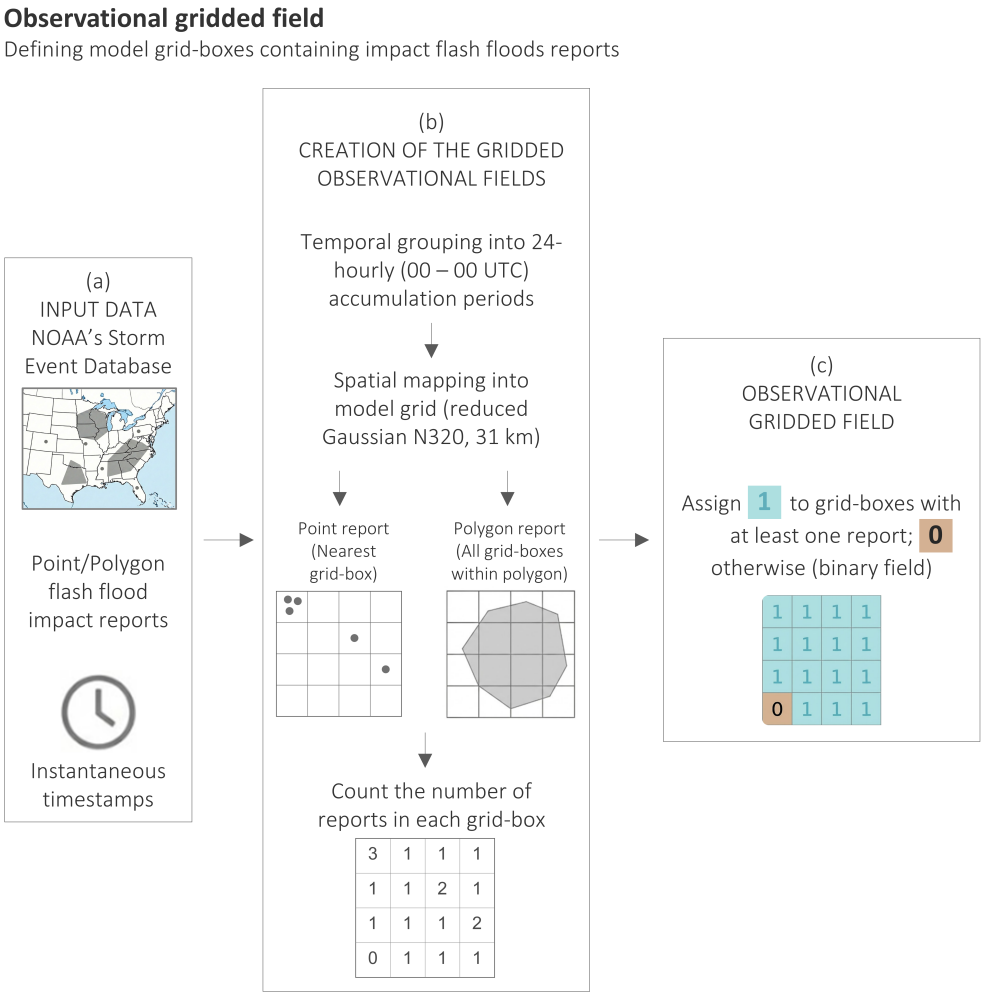

Figure 5.3: Schematic on how the gridded observational field is created from point/polygon impact reports with instantaneous timestamps. Panel (a) shows the input flash flood impact reports (from NOAA's Storm Event Database), consisting of point/polygon flash flood reports with instantaneous timestamps. Panel (b) shows the logic followed in the creation of the observational field, involving grouping reports into the considered accumulation period (24-hourly, 00–00 UTC) and spatially mapping them to the considered model grid (reduced Gaussian N320 at 31 km resolution at the equator). Point reports are assigned to the nearest grid-box, whereas polygon reports are assigned to all grid-boxes within the polygon. The total number of reports accumulated in each grid-box is counted. Panel (c) shows that the final observational gridded field is a binary classification, created by assigning a value of 1 to grid-boxes containing at least one report, and 0 otherwise.

5.2.2 Building the contingency table for probabilistic forecasts, using non-standard observations (impact reports)

Defining the contingency table for probabilistic forecasts

To assess the two desirable properties of probabilistic forecasts, i.e., reliability and discrimination ability (as discussed in Section 3.3 in Chapter 3), one must first construct the contingency table for probabilistic forecasts (Figure 5.4a). The construction proceeds in three steps. First, given all

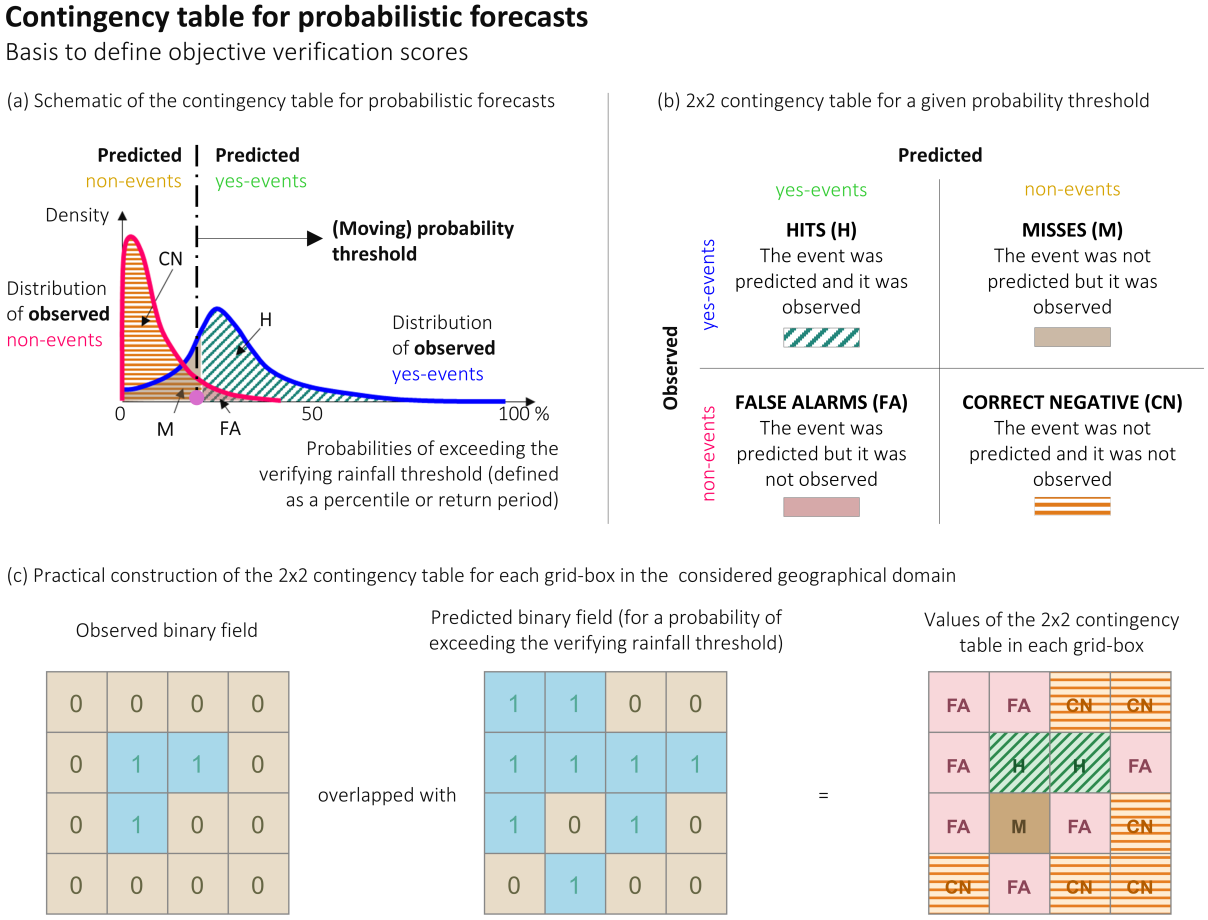

Figure 5.4: Contingency table for probabilistic forecasts. Panel (a) shows a schematic on how a contingency table for probabilistic forecasts is built. Panel (b) shows how, when fixing the probability of exceeding the verifying rainfall threshold, it is possible to build a 2x2 contingency table. A series of 2x2 contingency tables is obtained, one per threshold. Panel (c) shows a schematic of the practical construction of the 2x2 contingency table for each grid-box in the considered geographical domain.

cases of observed yes- and non-events (computed as described in Figure 5.3 in Section 5.2.1), one derives the distribution of observed yes- (blue distribution in Figure 5.4a) and that for observed non-events (distribution in pink). Second, a probability threshold for exceeding the verifying rainfall threshold (dotted-dashed vertical black line in Figure 5.4a) partitions the rainfall forecasts into predicted yes-events (to the right of the threshold) and predicted non-events (to the left). This partitioning yields four categories: hits (H, green diagonal hatching) — events correctly predicted and observed; false alarms (FA, pink shading) — events predicted but not observed; misses (M, brown shading) — events observed but not predicted; and correct negatives (CN, orange horizontal hatching) — non-events correctly predicted. Third, these counts populate a 2×2 contingency

table for that threshold (Figure 5.4b). In practice, the 2×2 contingency table is populated by comparing co-located grid boxes in the predicted and observed fields (built as defined in Figure 6.3 and 5.3). A hit occurs when corresponding grid-boxes are assigned a value of 1; a correct negative when both are 0; a miss occurs when the observed value is 1, but the forecast is 0, and a false alarm occurs when the forecast is 1, but the observed value is 0 (Figure 5.4c). By repeating this process across all considered probability thresholds, a series of 2x2 contingency tables is obtained, one per threshold.

Issues in the definition of the contingency table considering the non-stationarity of the observational dataset, and impacts on the verification analysis.

Unlike stationary observations (such as instruments installed at a specific location - e.g., rain gauges or discharge gauges - that provide a continuous timeseries of observed yes- *and* non-events), impact reports are non-stationary observations that record only observed yes-events. In the first case, all four quadrants of the contingency table can be quantified. Since in the latter case (when using impact reports as ground truth) it is impossible to answer the question ”if there are no reports at a location, is it because an event happened but nobody reported it, or because there was no event to report?”, it is more difficult to fill all four quadrants of the contingency table. Some studies using impact reports as ground truth verify only yes-events, with the caveat that only quadrant H (i.e. hits) and M (i.e. misses) of the contingency table can be populated (Robbins and Titley, 2018). This approach offers only a partial assessment of the forecasts' performance as it does not attempt to quantify FAs. In this thesis, the approach developed by Tsonevsky et al. (2018) and Pillosu et al. (2024) is adopted instead. They assume that a non-report represents an observed non-event. This assumption is acceptable for the considered impact reports, as they undergo an acceptable quality control. Nonetheless, given the constraints of the observational data, this approach will inherently inflate the number of FAs. However, this approach will provide a broader and more trustworthy evaluation of the rainfall-based predictions of areas at risk of flash floods.

5.2.3 Verification scores

As addressed in Section 3.3 in Chapter 3, the two properties of probabilistic forecasts should be assessed - reliability and discrimination ability. Moreover, it is also good practice to use overall scores - that integrate the forecasts' performance over the whole distribution of probability thresholds - and breakdown scores - that introduce granularity in the evaluation perfor-

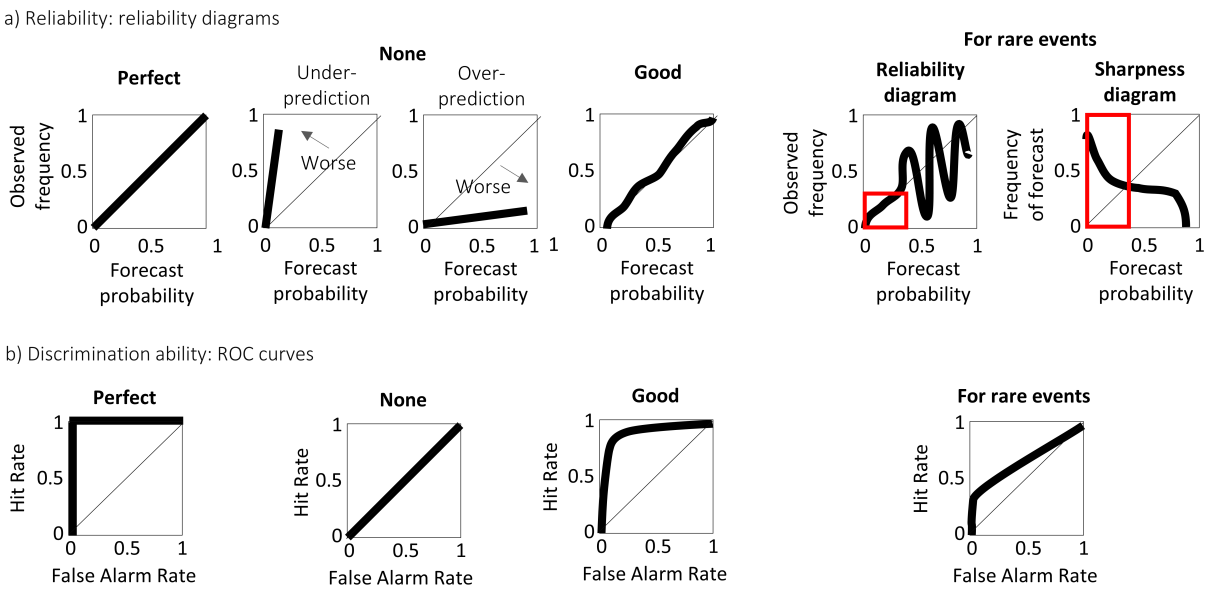

Figure 5.5: Breakdown verification scores. Panel (a) shows examples of reliability diagrams for forecasts with perfect, none, and acceptable reliability. An example of reliability diagram for the case of rare events is also shown, including the sharpness diagram. The red square indicates the forecast probabilities with the largest number of cases. Outside the red square, the reliability diagram becomes noisy. Panel (b) is similar, but for ROC curves.

mance, focusing the analysis on specific rainfall thresholds. The following sections describe the overall and breakdown scores considered to evaluate the forecasts' reliability and discrimination ability.

5.2.3.1 Reliability

Overall reliability: frequency bias

The frequency bias (FB) assesses the overall reliability of the rainfall-based predictions of areas at risk of flash floods. The frequency bias represents the fraction of the total number of predicted yes-events over the total number of yes-events in the observations. It is calculated with the equation 5.2:

$$\text{Frequency Bias} = \frac{H + FA}{H + M} \tag{5.2}$$

FB values range from 0 to $+\infty$, with FB = 1 indicating perfect bias. Values greater or smaller than 1 indicate, respectively, over- and under-prediction of the observed yes-events. It is worth noting that FB measure the overall ratio of forecast events to observed events and is not a measure of forecast

skill. As such, it can provide a score of 1 when there are compensating errors. Moreover, the FB might show large overestimations if the observed event is heavily underreported, as it is in our case.

Breakdown reliability: reliability diagrams

Reliability diagrams are used instead as breakdown scores to assess reliability (Figure 6.4a). They plot the relative forecast probabilities of an event against its corresponding relative observational frequency, indicating how reliable the forecast probabilities are at different probability classes. For perfect forecasts, when the forecasts show x% probability of occurrence, observations should meet the criteria x% of the time, so that the reliability curve lies on the diagonal. If the reliability diagram is above the diagonal for a specific forecast probability, those forecasts are under-predicting the likelihood of observing a yes-event. If it lies below the diagonal, there is over-prediction. When analysing reliability diagrams, especially when considering high verifying rainfall thresholds (i.e., rare events), it is important to know the frequency distribution of forecasts issued for specific probabilities (also called sharpness diagrams). For example, the small probability thresholds are the most important (within the red square in 6.4a). The sample of forecasts with high probabilities (outside the red square) will be rather small, and the reliability diagram is likely to appear noisy. Dimitriadis et al. (2021) propose a formulation for more stable reliability diagrams in case of rare events, but this formulation will not be considered in this study.

5.2.3.2 Discrimination ability

Breakdown discrimination ability: ROC curves

The Relative Operating Characteristic (ROC) curve is built from the probabilistic contingency table in Figure 5.4, mapping Hit Rates (HR) against False Alarm Rates (FAR), computed from equations 5.3 and 5.4, respectively:

$$\text{HR} = \frac{\text{H}}{\text{H} + \text{M}} \quad [\text{values between } 0 \text{ and } 1] \tag{5.3}$$

$$\text{FAR} = \frac{\text{FA}}{\text{FA} + \text{CN}} \quad [\text{values between } 0 \text{ and } 1] \tag{5.4}$$

HRs are mapped (Y-axis) against FARs (X-axis) in a unit square (Figure 6.4b). The form of the ROC curve illustrates how HRs vary with FARs as one systematically lowers the threshold probability at which it is assumed that an

event has been technically forecast to happen (i.e., from a 100% probability in the bottom left corner to a 0% probability at the top right corner).

Overall discrimination ability: area under the ROC (AUC-ROC)

The values of the geometrical area under the ROC curve (AUC-ROC) provide a summary measure of the discrimination ability across all probability thresholds. Perfect discrimination ability is obtained when only HRs grow, and FARs remain zero (Figure 6.4b). It is represented by an ROC curve that rises along the Y-axis from the bottom left corner of the unit square to the top left corner and moves straight to the top right corner. In this case, the AUC-ROC equals 1. If HRs and FARs grow at the same rate, the forecasts may appear to lack discrimination ability, as they perform similarly to a climatological forecast or due to a limited number of issued forecasts exceeding the verifying rainfall thresholds. In this case, the ROC curve lies along the diagonal, and AUC-ROC equals 0.5.

How the way AUC-ROC is computed can impact the interpretation of the verification results

How ROC curves and AUC-ROCs are computed can impact the interpretation of the forecasts' discrimination ability. For rainfall-based predictions, the ROC curves will be built for incremental decision thresholds that are materially assessable from the real ensemble configuration. In this way, we can estimate the ”real” forecast discrimination ability (Wilks, 2020). Probability thresholds are determined by considering the full discretisation ability in the ensemble (e.g., 99 members in the case of ERA5-ecPoint). This ensures that the ROC curves are as complete as possible (Bouallègue and Richardson, 2022). The number of thresholds corresponds, therefore, to the number of members exceeding the verifying rainfall threshold, so that for an ensemble of size M, maximum discretisation is achieved by M+1 probability thresholds (i.e., 0, 1/M, 2/M,..., M/M=1). The ROC curve is then built by straight segments joining successive points. It is then completed by joining that last meaningful point with a straight line in the top right corner of the unit square. For rare events (Figure 6.4b), the points of a ROC curve cluster in the graph's bottom left corner and completing the ROC with a straight line might give the impression that part of the ROC curve is missing (Casati et al., 2008). How much the curve appears incomplete depends on the ensemble size and the base rate of the event. The area under the ROC curve (AUC-ROC) will be computed using a trapezoidal approximation by adding the areas of single trapeziums formed by the straight lines between consecutive points in the ROC curve (Bouallègue and Richardson, 2022).

All scores are provided with 99% confidence intervals computed via bootstrapping with replacement over 1000 repetitions.

5.3 Results

5.3.1 Overall verification scores

Overall reliability: frequency bias (FB)

The frequency bias (Figure 5.6a) reveals systematic overprediction across all return periods, with values relatively stable across all lead times. The frequency bias is larger - around 70 - for less severe rainfall events (e.g., 1-year return period). When considering a slightly more severe event (e.g., 5-year return period), the frequency reduces to a third - around 20, and it becomes remarkably smaller - around 2 - for very extreme rainfall events (e.g. 50-year and 100-year return period, see inset in Figure 5.6a).

Overall discrimination ability: area under the ROC curve (AUC-AUC-ROC)

The area under the ROC curve (AUC-AUC-ROC) shows a good overall discrimination ability (always above 0.5) across all return periods. Values for the rainfall reanalysis (lead time for day 0) are similar to those for the forecasts at day 1, and then they slightly decrease from day 2. AUC-ROC Values for rainfall exceeding 1-year return period are around 0.7 at day 0 and 1, and they do not go under 0.6. AUC-ROC Values for rainfall exceeding 100-year return period (very extreme events) do not vary much over lead time, but remain just above 0.5.

5.3.2 Breakdown verification scores

Breakdown discrimination ability: ROC curve

All forecasts for rainfall events exceeding the 1-year return period threshold (Figure 5.7) exhibit a discrimination ability superior to random chance, as the curves are above the diagonal reference line. A systematic degradation in discrimination ability is observed with increasing lead time, with the Area Under the ROC Curve (AUC-ROC) values ranging from 0.675 for the short-range forecasts (Figure 5.7a) to 0.612 ($\sim$9% reduction) for t+120 (day 5, Figure 5.7f). Despite such a reduction, the forecasts show a good discrimination ability throughout the forecast horizon. Day 1 forecasts (t+24) show a higher discrimination ability than the short-range forecasts, and only from day 2 forecasts (t+48), the discrimination ability of the long-range forecasts goes below that of the reanalysis. The relatively narrow confidence intervals (at 99% confidence level) suggest that the differences in skill between forecast configurations are statistically meaningful at the considered confidence level. The black square in Figure 5.7a shows that perfect reliability (i.e. frequency bias equal to 1) is reach for probabilities <= 23%. For the long-range forecasts, the probability thresholds at which perfect reliability is achieved is compatible to the short-range, being 27% for

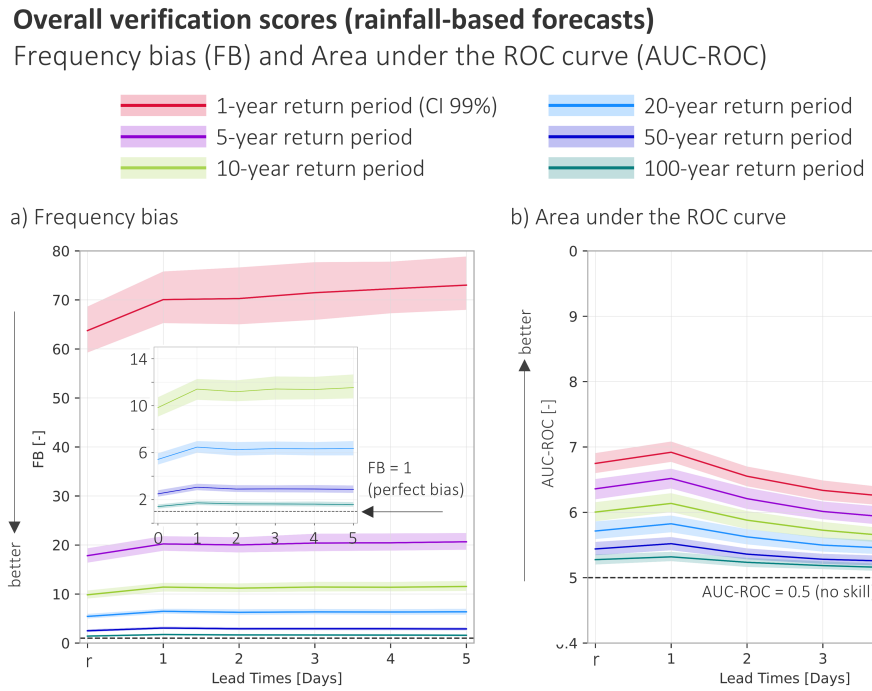

Figure 5.6: Overall verification scores for the rainfall-based forecasts of areas at risk of flash flood. Panel (a) shows the frequency bias (solid lines) for 1-year (in red), 5-year (in purple), 10-year (in light green), 20-year (in cyan), 50-year (in blue), and 100-year return period (in green). The corresponding shaded areas represent the confidence intervals at 99% confidence level. The inset box contains a zoomed-in version of the panel to better show the frequency bias values close to 1 (representing perfect bias). Panel (b) shows the area under the ROC curve. Lead time equal to r" relates to the statistics computed for ERA5-ecPoint *reanalysis*, while lead times from 1 to 5 (in days) relate to ERA5-ecPoint *forecasts*.

all the lead times except t+120 which is 26% (Figure 5.7b-f). The frequency bias for the lowest probability threshold (i.e. 99th percentile or probability threshold equal to 1%) in the short-range forecasts equals to 28. The frequency biases for the long-range forecasts are similar, falling between 31 and 33. Similar results are obtained for the 5-, 10-, 20-, 50-, and 100-year return periods.

Breakdown reliability: reliability diagrams

All forecasts for rainfall events exceeding the 1-year return period threshold (Figure 5.9) exhibit a systematic overprediction across all lead times, as shown by the reliability diagram being below the diagonal line. This indicates that when the model predicts a given probability, the observed frequency of flash flood events is consistently lower. For example, when the forecasts indicate a 50% chance of having a flash flood event, the observed frequency ranges from ~10% in the short-range forecasts (Figure 5.9a), and between 10% (for t+24, Figure 5.9b) and 2% (t+120, Figure 5.9f) in the long-range forecasts. As seen in the ROC curves, the confidence intervals at 99% are fairly narrow, suggesting fairly confident estimates

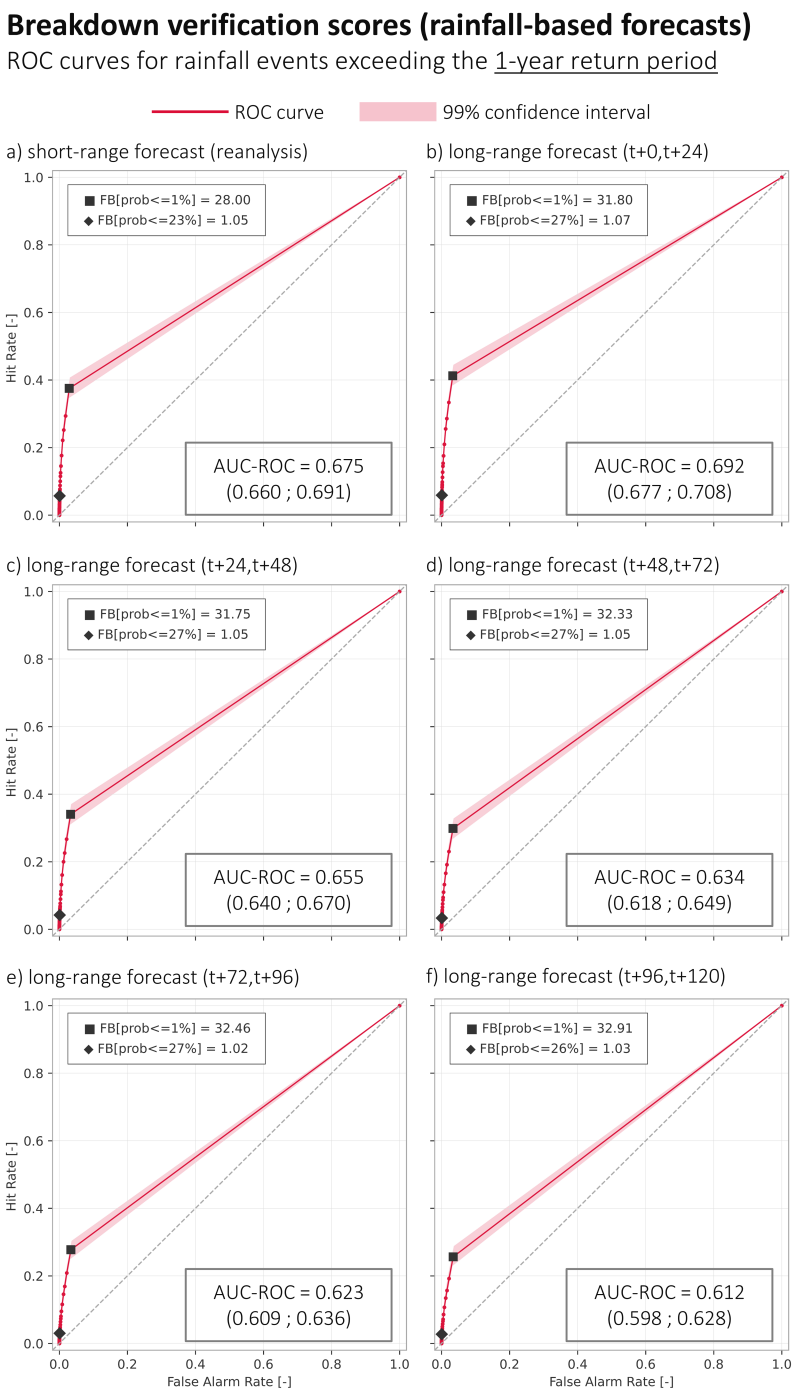

Figure 5.7: ROC curves for tp >= 1-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint. Panel (a) shows the ROC curve (red solid line) for the ERA5-ecPoint reanalysis together with the confidence intervals (red shaded area) at 99% confidence level. Panels (b) to (f) refer to ERA5-ecPoint forecasts, for accumulation periods ending in t+24, t+48, t+72, t+96, and t+120, respectively. The dots with the *diamond* symbol refer to the probability threshold at which the frequency bias has the closest value to 1 (i.e., perfectly reliable forecast), while the dots with the *square* symbol show the value of the frequency bias for the lowest probability threshold available in ERA5-ecPoint (i.e., the 99th percentile).

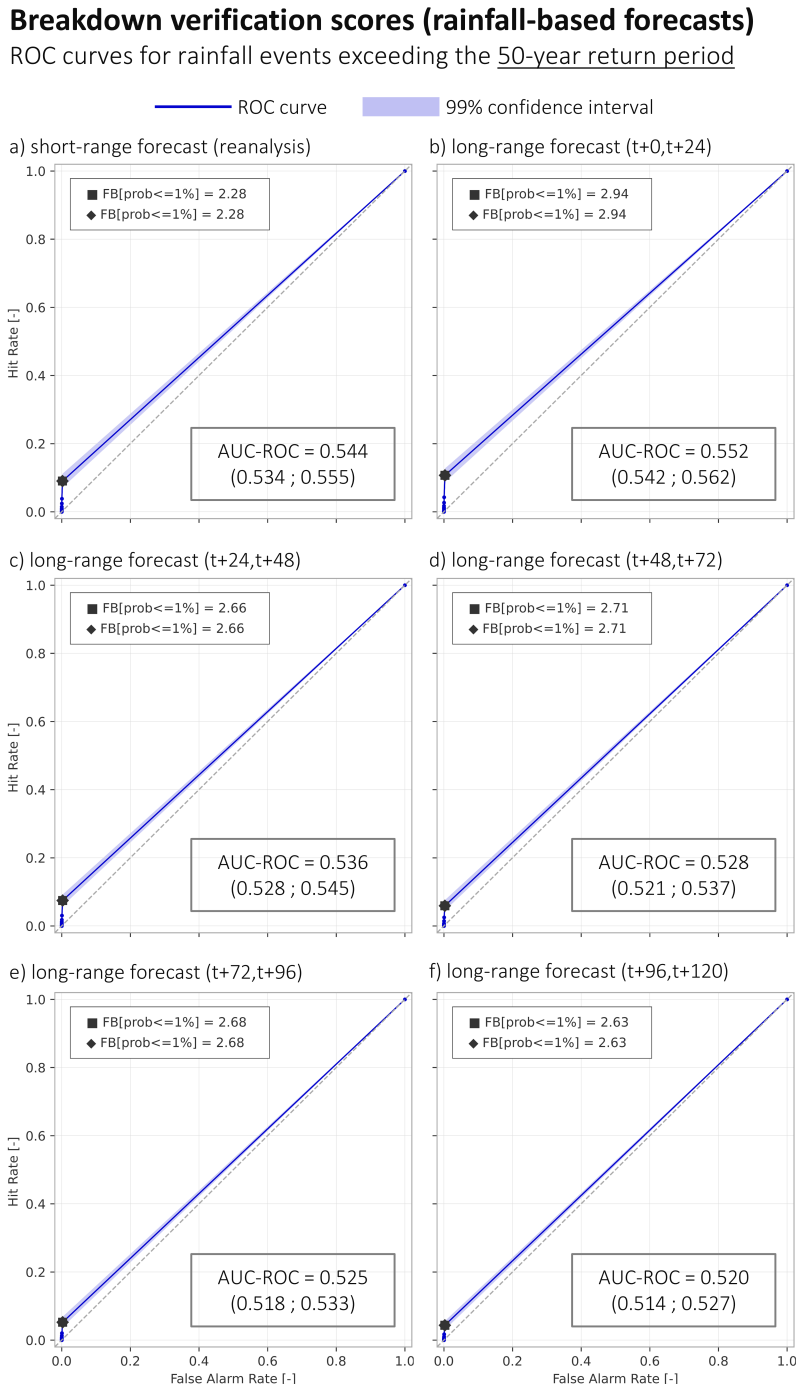

Figure 5.8: ROC curves for tp >= 50-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint. Similar to Figure 5.7.

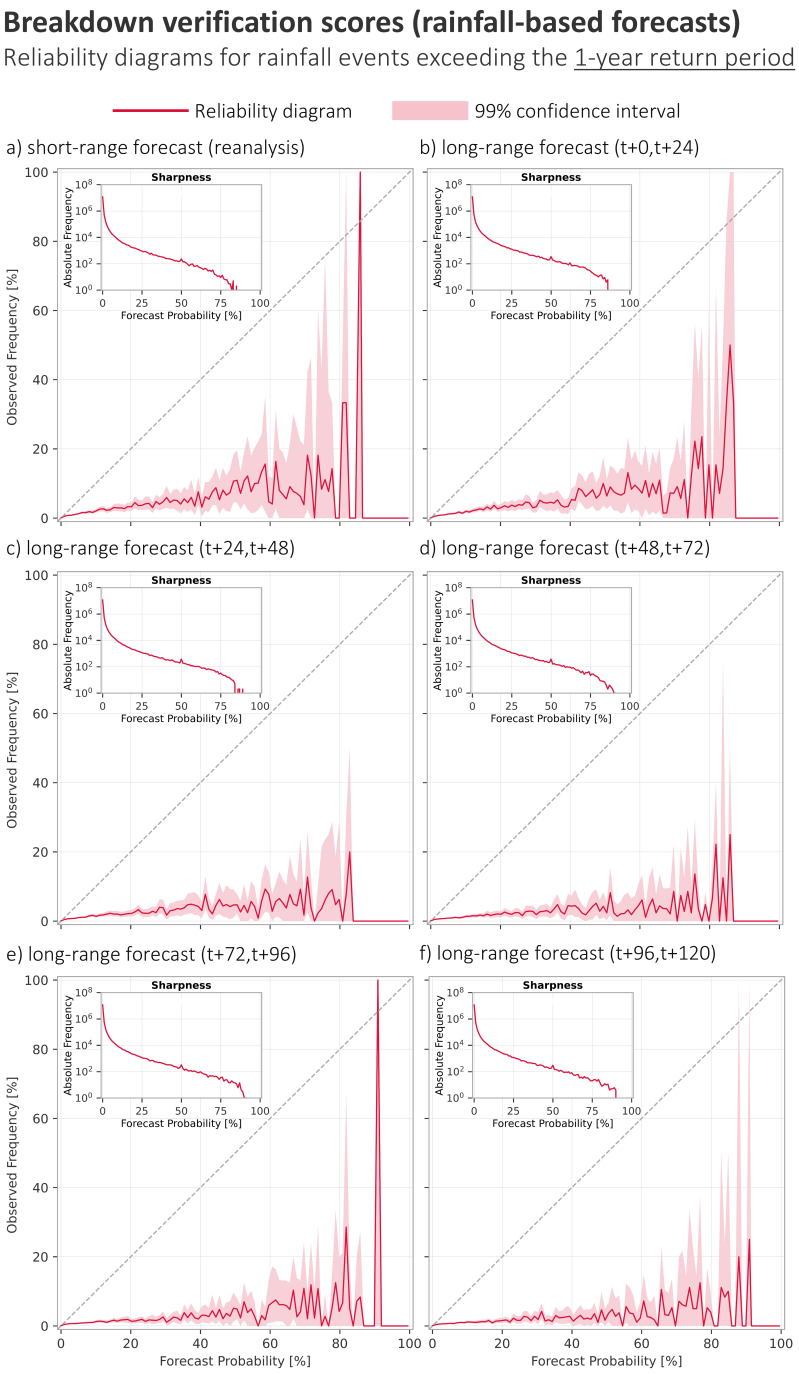

Figure 5.9: Reliability diagrams for tp >= 1-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint. Panel (a) shows the reliability diagram (red solid line) for the short-range predictions together with the confidence intervals (red shaded area) at 99% confidence level. Panels (b) to (f) refer to the long-range forecasts for accumulation periods ending in t+24, t+48, t+72, t+96, and t+120, respectively. The inset boxes show the corresponding sharpness diagrams.

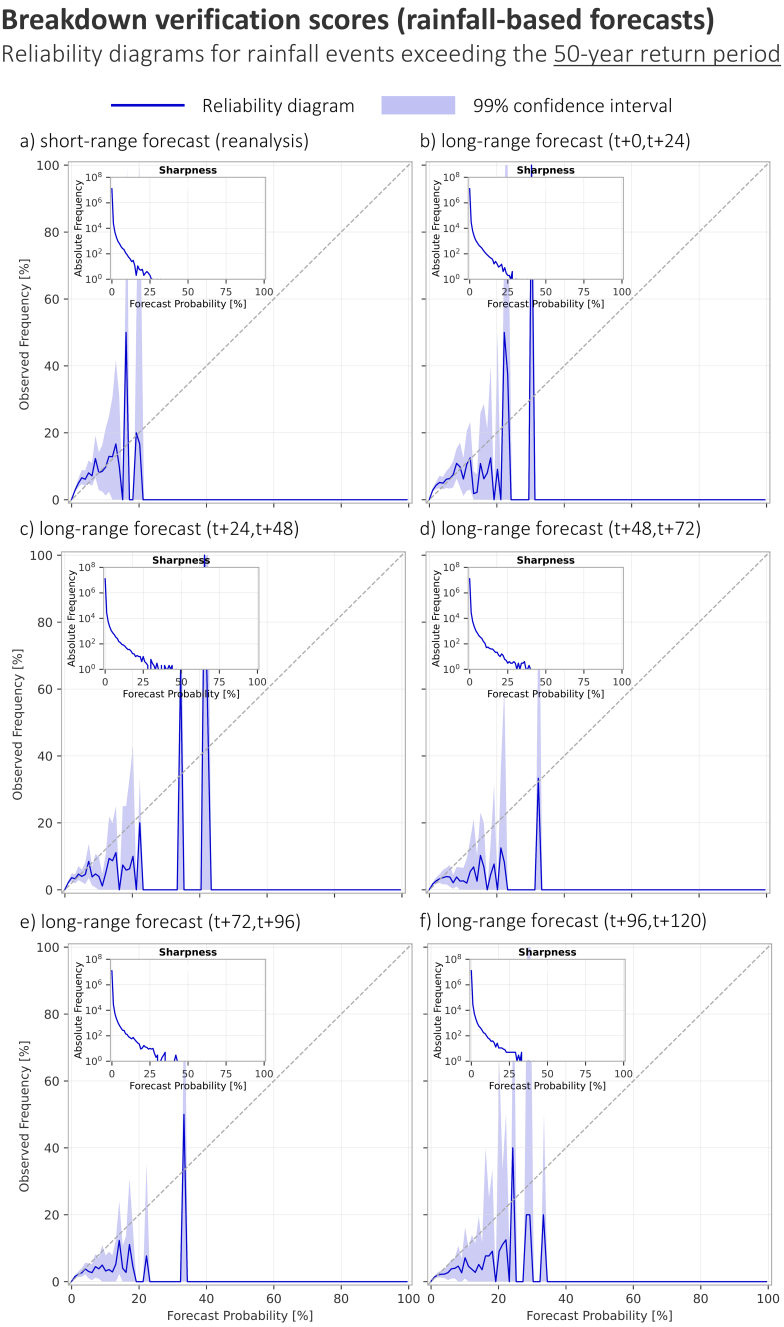

Figure 5.10: Reliability diagrams for tp >= 50-year return period for the rainfall-based forecasts of areas at risk of flash floods built with ERA5-ecPoint. Similar to Figure 5.9.

of the reliability diagrams for forecast probabilities less than ~25%. The confidence levels increase with increasing forecast probabilities due to the low number of forecasts issued with probabilities higher than 25%, when the total number of instances lies below 1000 samples, as seen in the

corresponding sharpness diagrams (inset boxes in all panels of Figure 5.9). A notable characteristic of all the reliability diagrams in Figure 5.9 is the sharp increase in observed frequency for the highest probability bins (i.e. 80% to 100%). This steep rise suggests that when the model does issue high probability forecasts, these correspond to genuinely extreme events, though such forecasts remain infrequent. The temporal evolution from Figure 5.9a-f reveals subtle changes in the forecasts' reliability characteristics with increasing lead time. Whilst the general pattern of overprediction persists, from day 2 forecasts (t+48) results more squashed into a flat line over very small observed frequencies, indicating that even though the model issues forecasts with high probabilities of exceeding the 1-year return period at longer lead times with a similar frequency of the short-range forecasts and the day 1 (t+24), such forecasts do not necessarily correspond to an observed flash flood event. Finally, the reliability diagrams get closer to the diagonal line (indicating perfect bias) as we increase the rainfall threshold to identify the flash flood events. Reliability improves for small probabilities (typically below 10%, at all lead times) as the return period increases (the 50-year return period, in Figure 5.10a-f is shown as an example).

5.4 Case study on Storm Ida

Storm Ida case study: strong predictive signal over New York maintained to day 5; weaker, scattered signal for western convective events

The objective verification results can be further interpreted by analysing the probabilities of exceeding the 1- and 50-year return period for the case study on Storm Ida[3]. Over the New York area, the probabilities of rainfall exceeding the 1-year return period are very high - >80% - in ERA5-ecPoint reanalysis (Figure 4.7a, blue circle). It also indicates overall very well the general area impacted by Storm Ida as seen in the observations in Figure 4.2d. The overall signal of potential extreme rainfall is given by the probabilities of exceeding the 50-year return period also being fairly high - >10% - in ERA5-ecPoint reanalysis (Figure 4.8a). The signal of extreme rainfall is well maintained up to day 5 lead time (Figures 4.7b-f and Figures 4.8b-f, blue circles). The picture is different for the extreme rainfall events on the Western side. The areas potentially affected by moderate rainfall (Figure 4.7a, red circle) are much larger than those observed in Figure 4.2d. Indeed, the probabilities of exceeding the 50-year return period

[3]For a detailed description of the synoptic conditions during Storm Ida and its impacts, please refer to Section 4.5 in Chapter 4.

are much smaller, not exceeding 5%. Moreover, it is worth noting the *scattered pattern* of the areas highlighted by the rainfall predictions and the low predictability at longer lead times (Figures 4.7b-f and Figures 4.8b-f, red circles). These patterns are more commonly observed during small-scale convective systems, which tend to drive very localised flash flood events.

5.5 Discussion

Reliability and discrimination ability of the rainfall-based predictions of areas at risk of flash floods

For high-frequency events (e.g., 1-year return period), the high-frequency bias values indicate that ERA5-ecPoint forecasts identify potential flash flood areas up to 70 times more frequently than such events are observed in the Storm Event Database. In marked contrast, extreme events (with return periods of 50-100 years) demonstrate near-optimal frequency bias values between 2 and 3. These results may have two interpretations. The first one concerns the quality of the forecasts, indicating that rainfall-based predictions of areas at risk of flash floods based on rainfall forecasts exceeding at least the 50-year return period are the ones that should be considered for the prediction of areas at risk of flash floods, as they are the ones providing near-optimal reliability. The second interpretation concerns the type of flash flood events recorded in the database, i.e., mostly extreme flash flood events, generated by rainfall events exceeding the 50-year return period, are recorded in the database. It is not possible to disentangle from the information at hand which of the two interpretations might be the correct one. The case study corroborates this result. Forecasts of rainfall exceeding the 1-year return period might indicate a far too big area at risk of flash flood in the case of small-scale convective systems, which tend to generate more localised flash flood events. Since there may not be as many flash flood events, or they may occur in areas that go underreported, these low-return period forecasts may yield very high frequency biases. Hence, it is advocated that the objective verification framework should at least distinguish between flash floods generated by small-scale convective systems and larger-scale systems because the NWP rainfall forecasts may yield different performance levels in these two cases. This hypothesis aligns with results in other studies (Pillosu et al., 2024).

Interpreting verification metrics with sparse observational coverage

Although the Storm Event Database correspond to one of the best attempts to build a comprehensive historical record of flash flood events (in the US), event underreporting still needs to be addressed to provide a more comprehensive assessment of forecast performance. In alignment with

previous studies attempting to use impact-based observations to estimate forecast performance (Hitchens et al., 2013; Robbins and Titley, 2018; Mitheu et al., 2023, 2025), the verification of the rainfall forecasts against underreported flash flood events can lead to an underestimation of forecast skill and undermine the confidence in the forecasts, causing the dismissal of valuable predictions crucial for preparedness actions. It is, therefore, required to read beyond the actual numbers of verification results, and read them in a critical, although subjective, way. For example, although the FB is far larger than 1 (that would mean the forecasts overestimate substantially the areas at risk of flash floods), areas of yes-events in the forecast show a good correspondence with areas having some or at least one flash flood report. This shows that the forecasts are at least able to provide guidance on the general location of areas at risk of flash floods, and that the high frequency bias is due to the bigger number of grid-boxes in the forecasts with a yes-event compared to the lower number of flash flood reports. While this lower number of flash flood reports might represent a genuine forecast overestimation of the areas at risk of flash flood, such lower number might be due as well to the fact that a single flash flood report is assigned to a single point (and consequently a single grid-box), but it perhaps represents a bigger area, or more flash flood events were unreported. Techniques such as assigning 1s (i.e., yes-events) to adjacent grid-boxes (buffer area) to those already containing flood reports could help to reduce the FB. We argue, however, that assigning 1s to adjacent grid-boxes is more appropriate for forecasts provided on higher resolution grids because rainfall might not be predicted at the right location, flash flood events might extend to adjacent grid-boxes, or the uncertainty in the reporting locations might cover more grid-boxes. Since forecasts are provided on a grid at 31 km, if a flash-flood-triggering rainfall event is predicted within a grid-box, it is reasonable to think that there is a good chance of seeing the flash flood within that grid-box (unless the event happens on the grid-box boundary, but this is an exception whose handling goes beyond the scope of this analysis).

Adequacy of the impact observations for verifying flash floods

Event though the scientific community must strive in continuingly improving global NWP rainfall forecasts, it is essential to improve the spatial coverage of flash flood impact reports to show forecast users a number of false alarms closer to that related to forecast skill rather than that mainly due to underreporting. For this type of analysis (i.e., verification of forecasts using non-standard observations) is the biggest and most difficult problem to address (Marsigli et al., 2021). This study demonstrates that enhanced flash flood report databases are instrumental in assessing the performance

of rainfall forecasts for flash flood prediction. The improved and more spatio-temporal coverage of flash flood reports in the Storm Event Database enabled an in-depth, long-term assessment that would have been unfeasible with other databases, such as EM-DAT, due to their poorer spatio-temporal coverage. For these reasons, flash flood verification in the past was primarily based on case studies or at catchment level, as more detailed information is available for single events (Gaume et al., 2009, 2016). While a case-study-based verification approach or at catchment level is invaluable in understanding how forecasts predict flash flood events, provided enough observations are available, the results may not hold for other events due to the focused nature of the analysis. Alternatively, by leveraging the higher quality and spatial coverage of rainfall observations, researchers can use them to assess the performance of rainfall forecasts in predicting flash floods. However, as seen in this study, the results of these two verification analyses are different. ecPoint almost always performs better than ENS in predicting extreme (localised) rainfall (Gascón et al., 2024; Hemri et al., 2022; Hewson and Pillosu, 2021). However, in the binary prediction of whether an area was affected by a flash flood or not (yes- or non-event), the verification results are more nuanced. Failing to consider these two cases separately would do a disservice to the assessment of how well raw ENS forecasts can predict areas at risk of flash floods. Thus, this study highlights the importance of enhancing flash flood report databases to assess the performance of rainfall forecasts for flash flood prediction in more detail, thereby contributing significantly to more effective disaster preparedness and risk management strategies.

5.6 Conclusion

Development of the first objective verification framework for evaluating (rainfall-based) predictions of areas at risk of flash floods

This chapter presents the first systematic flash-flood-focused verification framework for assessing global NWP rainfall forecasts in the binary identification of areas at risk of flash floods. The framework, while developed using ERA5-ecPoint rainfall forecasts verified against NOAA's Storm Event Database over CONUS, is sufficiently flexible to accommodate any gridded rainfall forecast product, thereby enabling model intercomparison and providing a benchmark against which more sophisticated approaches (e.g., incorporating hydrological or topographical parameters) can be measured. Such comparative analysis is essential for determining whether increased model complexity yields commensurate improvements in prediction accu-

racy or whether precipitation-based approaches provide sufficient utility for early warning applications.

Rainfall thresholds with higher return periods are recommended for operational flash flood guidance

The verification results demonstrate that overall discrimination ability, measured by AUC-ROC, and overall reliability, measured by frequency bias, remain relatively stable across lead times from day 1 to day 5. Notably, day 1 forecast performance is comparable to that of the reanalysis, with degradation becoming apparent only from day 2 onwards. Moreover, the ROC curves reveal that the post-processed ERA5 reanalysis and forecasts identify observed yes-events with minimal false alarms across all lead times for cases of extreme rainfall (return periods of 50-years), and reliability diagrams indicate near-perfect reliability for these extreme thresholds at useful probability thresholds. This pattern suggests that rainfall-based predictions employing higher return period thresholds may be more appropriate for operational flash flood guidance, though the extent to which this reflects genuine forecast behaviour versus database characteristics remains difficult to disentangle. The Storm Ida case study further illustrates that forecasts of rainfall exceeding lower return periods may indicate excessively large areas at risk during small-scale convective events, reinforcing the need to distinguish verification results by synoptic regime.

The critical role of flash flood impact reports' quality and quantity to model verification and development

This study also underscores the critical role of observation quality in forecast verification. Although the Storm Event Database represents one of the most comprehensive flash flood impact archives available, substantial underreporting persists and inevitably inflates false alarm rates. The enhanced spatio-temporal coverage of this database, however, enabled an in-depth, multi-year verification analysis that would have been infeasible with sparser alternatives such as EM-DAT. Hence, continued investment in developing and maintaining high-quality flash flood databases remains essential. Such databases provide invaluable information on flood typology that can inform targeted verification efforts and guide forecast development. Until reporting coverage improves substantially, verification metrics should be interpreted with appropriate caution, recognising that elevated frequency bias values may reflect observational gaps as much as forecast deficiencies.

CHAPTER 6

DATA-DRIVEN HYDRO-METEOROLOGICAL PREDICTIONS OF AREAS AT RISK OF FLASH FLOOD: FROM SHORT- TO MEDIUM-RANGE LEAD TIMES

6.1 Introduction

Opportunities for data-driven flash flood prediction at continental scale

Chapters 1 and 2 of this thesis have established that extending flash flood predictions to medium-range lead times (i.e., from day 1 to 5) over large spatial domains remains an unresolved challenge in modern hydrology. Yet recent scientific advances suggest a viable path forward. First, Chapter 5 demonstrated that statistically post-processed global NWP rainfall forecasts can identify areas at risk of flash floods with good reliability and discrimination ability, up to day 5. These predictions, however, rely solely on rainfall information. Catchment response to heavy rainfall is governed by hydrological characteristics, e.g., antecedent soil moisture, topographic steepness, and land surface conditions (Zanchetta and Coulibaly, 2020). This interaction has traditionally been explained with physically-based hydrological models (Panigrahi et al., 2025). However, running them over a continuous global domain and up to medium-range lead times has been technically and economically prohibitive (Philipp et al., 2016). Data-driven approaches offer a computationally efficient alternative that can leverage the knowledge accumulated through decades of physical modelling at

catchment scale and index-based model development at continental scale (Kratzert et al., 2019).

Chapter objectives: assessing the feasibility of data-driven, medium-range predictions of areas at risk of flash floods, and predictability horizon

This chapter evaluates the feasibility of developing data-driven predictions of flash flood risk at continental scale, addressing Research Question 2 of this thesis (see Section 1.1). The primary objective is to determine whether data-driven architectures can produce skilful probabilistic predictions of areas at risk of flash floods, using ERA5 reanalysis and forecasts to build the model's hydro-meteorological features[1] and regional flash flood impact reports as the model's target variable[2]. These choices reflect deliberate design decisions to enable future global scalability (which will be explored in Chapter 7). ERA5, despite its coarse resolution for flash flood applications (31 km), provides continuous global coverage. Moreover, impact reports may provide broader global observational coverage than discharge gauge networks, given advances in AI-based extraction of impact information from global news and social media (Robbins and Titley, 2018; Spruce et al., 2021; Wyatt et al., 2023), making it timely to evaluate their suitability for supervised learning. A secondary objective is to establish the predictability horizon of the data-driven predictions, i.e., the lead time beyond which the forecast skill degrades below operationally useful thresholds. The evaluation will consider the reanalysis-based estimates as the upper bound of achievable skill, alongside medium-range forecasts representing realistic operational conditions.

Methodological framework: model selection, training approaches to address class imbalance, and interpretability

Six distinct architectures are evaluated (comprising random forest and gradient boosting implementations, alongside feed-forward neural networks) to identify which model complexity is warranted for this application and whether simpler, more interpretable models can match the performance of more complex alternatives (Merz et al., 2022). The methodological framework addresses two challenges inherent to rare event prediction under class imbalance conditions (Altalhan et al., 2025). First, nested stratified cross-validation mitigates overfitting during hyperparameter tuning, a risk that

[1]The word "feature" indicates the input variables (e.g. probabilities of rainfall forecasts exceeding a certain return period, orographic characteristics, or antecedent soil moisture) used by a data-driven model to make predictions" after the first mention of "features". I've included domain-specific examples relevant to your hydro-meteorological context; adjust these if you prefer different examples.

[2]In supervised learning, the target variable is the known outcome used to train the model to make predictions on new data.

is particularly acute in severely imbalanced problems. Second, Bayesian optimisation systematically compares balanced versus weighted loss functions to determine the optimal strategy for handling class imbalance, with direct implications for the trade-off between detection rate and false alarm rate. Lastly, SHAP values provide physical interpretability by revealing how rainfall and hydrological parameters combine to produce the predictions, enabling diagnosis of potential spurious correlations (Rozemberczki et al., 2022). Finally, the verification framework employs metrics designed for imbalanced classification, such as the Precision-Recall curve (Sofaer et al., 2019), complementing the rainfall-focused verification presented in Chapter 5.

Chapter outline

The remainder of this chapter is organised as follows. Section 6.2 details the datasets used, the machine learning architectures, training procedures, and verification framework considered. Section 6.3 presents the results, covering model training diagnostics, verification over reanalysis data, and verification over medium-range forecasts. Section 6.4 examines the physical interpretation of model behaviour through SHAP analysis. Section 6.5 applies the selected model to Hurricane Ida as a case study. Section 6.6 discusses the results' implications, while Section 6.7 draws the conclusions of the study.

6.2 Data and methods

6.2.1 Feature and target variables

Description of the feature and target variables

The data used in this chapter have been presented in Chapter 4. The target variable, i.e., flash flood impact reports from the Storm Event Database, is described in Section 4.2. The hydrological (antecedent soil moisture), climatological (vegetation coverage), and static (orography slope) feature variables are described in Section 4.3. Finally, the post-processed rainfall reanalysis and forecasts, used for the development of the data-driven models, are described in Section 4.4.

6.2.2 Development of data-driven models

6.2.2.1 Model architecture selection

Six distinct machine learning architectures were evaluated to identify the optimal approach for flash flood probability estimation: *random forest*

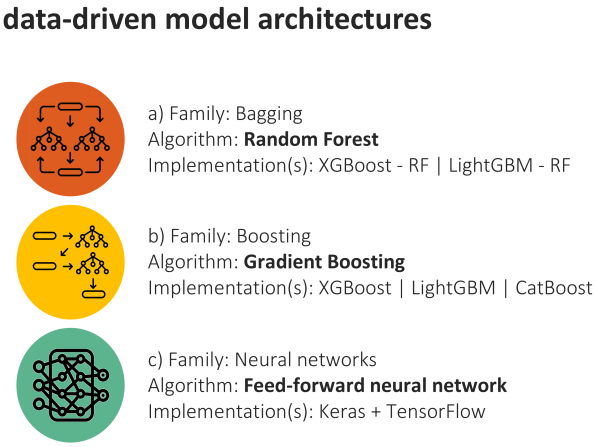

Figure 6.1: Overview of the three ensemble data-driven model architectures used in this study. *Bagging*, with random forests (with XGBoost and LightGBM implementations in random forest mode), *boosting*, with gradient boosting (with XGBoost, LightGBM, and CatBoost implementations in gradient boosting mode), and *neural networks*, with feed-forward architectures (implemented using Keras with TensorFlow backend).

(with XGBoost and LightGBM implementations), *gradient boosting* (with XGBoost, LightGBM, and CatBoost implementations), and a *feed-forward neural network* (constructed using Keras and TensorFlow). These models were selected based on their proven efficacy in handling tabular data with class imbalance and their ability to capture complex non-linear relationships between hydro-meteorological predictors and flash flood occurrence (Shwartz-Ziv and Armon, 2022).

Model architecture: random forest (XGBoost and LightGBM implementations)

Random forest (Figure 6.1a) constitutes an ensemble learning method that constructs multiple decision trees during training and outputs predictions based on the mode of individual tree classifications for categorical targets or mean predictions for regression tasks (Liu et al., 2012). The algorithm introduces randomness through two primary mechanisms: bootstrap aggregating (bagging), where each tree is trained on a random sample of the data with replacement, and random feature selection at each split, where only a subset of features is considered for determining the optimal partition. This dual randomisation strategy reduces overfitting and improves generalisation performance, particularly for high-dimensional datasets with complex feature interactions. The XGBoost implementation of random forest[3] adapts the gradient boosting framework to emulate a random forest behaviour by setting specific hyperparameters. This implementation maintains the

[3]https://xgboost.readthedocs.io/en/stable/tutorials/rf.html

core Random Forest principles whilst leveraging XGBoost's computational efficiency and regularisation capabilities. The key modifications include setting the number of parallel trees equal to the number of estimators, using a learning rate of 1.0, and disabling boosting rounds. This approach benefits from XGBoost's optimised handling of missing values, built-in cross-validation support, and efficient memory usage through its column block structure. The LightGBM implementation of random forest utilises gradient-based one-sided sampling and exclusive feature bundling techniques, which significantly reduce computational complexity, whilst maintaining accuracy[4]. The histogram-based algorithm constructs decision trees by bucketing continuous features into discrete bins, enabling faster split finding and reduced memory consumption. This implementation particularly excels with large-scale datasets and high-dimensional feature spaces.

Model architecture: gradient boosting (XGBoost, LightGBM, and CatBoost implementations)

Gradient boosting (Figure 6.1b) represents a powerful ensemble technique that sequentially constructs decision trees, where each subsequent tree attempts to correct the residual errors of the preceding ensemble (**?**). The algorithm operates by fitting new models to the negative gradients of a differentiable loss function, effectively performing gradient descent in function space. This iterative refinement process enables the capture of complex nonlinear relationships and interactions between features. The mathematical formulation frames boosting as an optimisation problem in function space, where the objective is to minimise the expected value of a loss function by iteratively adding weak learners. The use of shallow trees as base learners provides regularisation through structural constraints, whilst shrinkage parameters control the contribution of each tree to prevent overfitting. XGBoost (Extreme Gradient Boosting) enhances the traditional gradient boosting algorithm through several algorithmic innovations. The framework incorporates a second-order Taylor expansion of the loss function, enabling more accurate optimisation steps. Regularisation terms in the objective function penalise model complexity through both L1 and L2 norms on leaf weights and the number of leaves. The column block structure for parallel processing, cache-aware access patterns, and out-of-core computation capabilities enable efficient training on large datasets. For imbalanced classification problems, XGBoost provides the *scale_pos_weight* parameter to adjust for class frequencies directly in the loss function (Chen and Guestrin, 2016). LightGBM (Light Gradient Boosting Machine) introduces

[4]https://lightgbm.readthedocs.io/en/latest/index.html

novel techniques to accelerate training whilst maintaining accuracy. The Gradient-based One-Side Sampling (GOSS) retains instances with large gradients whilst randomly sampling from instances with small gradients, effectively focusing computational resources on difficult-to-classify examples. Exclusive Feature Bundling (EFB) reduces dimensionality by bundling mutually exclusive features, particularly beneficial for sparse datasets. The histogram-based algorithm and leaf-wise tree growth strategy enable faster convergence and better accuracy compared to level-wise approaches, though requiring careful regularisation to prevent overfitting (Ke et al., 2017). CatBoost addresses several fundamental challenges in gradient boosting through innovative algorithmic solutions. The ordered boosting approach mitigates prediction shift—a subtle form of overfitting in gradient boosting—by using different data permutations for calculating gradients and applying models. This technique provides unbiased gradient estimates and improves generalisation. The algorithm's symmetric tree structure, whilst potentially less flexible than asymmetric trees, enables extremely fast inference and natural handling of categorical features through novel encoding schemes. For imbalanced datasets, CatBoost offers sophisticated class weighting mechanisms and custom loss functions optimised for rare event detection (Prokhorenkova et al., 2018).

Model architecture: feed-forward neural networks: TensorFlow and Keras implementation

Feed-forward neural networks (Figure 6.1c) represent the fundamental architecture of deep learning, comprising layers of interconnected nodes where information flows unidirectionally from input to output. Each neuron computes a weighted sum of its inputs, applies a non-linear activation function, and propagates the result forward. This architecture's universal approximation capability—the theoretical ability to approximate any continuous function given sufficient neurons—provides the flexibility to model complex non-linear relationships in hydro-meteorological applications. The multi-layer perceptron architecture employed in this study consists of fully connected layers with rectified linear unit (ReLU) activations, providing non-linearity whilst mitigating gradient vanishing issues. Dropout regularisation randomly deactivates neurons during training, creating an implicit ensemble effect that reduces overfitting. The back-propagation algorithm, combined with adaptive learning rate optimisation through Adam (Adaptive Moment Estimation, an algorithm that adjusts learning rates for each parameter based on estimates of first and second moments of the gradients), enables efficient training even with limited positive examples in imbalanced datasets. The Keras implementation provides a high-level interface to TensorFlow's computational graph framework, enabling rapid prototyping whilst maintain-

ing computational efficiency. The Sequential — a linear stack interface that allows layers to be added one at a time in sequence — facilitates straightforward construction of feed-forward architectures with customisable depth and width. The implementation leverages TensorFlow's automatic differentiation capabilities for gradient computation and supports various optimisation algorithms, loss functions, and regularisation techniques. For imbalanced classification, the class_weight parameter in the fit method enables sample weighting to address class frequency disparities.

6.2.2.2 Feature engineering

Feature engineering: to address convective cell movement and water routing

The data-driven models used the raw variables primarily as described in Section **??**. The only variable engineered corresponds to the *maximum probability of 24-hourly rainfall exceeding a specific threshold in adjacent grid-boxes*. This variable examines the probabilities in adjacent grid-boxes to that of interest, within an assigned radius, and selects the maximum value. This variable addresses two critical limitations that arise when the identification of areas at risk of flash floods relies solely on the probability of exceeding a certain threshold over the grid-box of interest. The first (meteorological) reason relates to the convective parametrisation scheme in global NWP models. In global NWP models, convective cells do not move, meaning the rainfall falls where the convective cell was generated by the model (Doswell, 2001). In reality, convective cells move in the direction of the wind (Doswell, 2001). A typical case corresponding to this scenario is a more or less organised convective system generated over a warm water body (e.g., the Mediterranean) that then moves onto land. Such a convective system, if conditions are favourable, may deliver significant rainfall amounts over land. However, as far as the model is concerned, the rainfall may fall over the water body, potentially causing a large underestimation of rainfall estimates over land. The second (hydrological) reason concerns the absence of water routing (over land or water courses) in this analysis. When rainfall occurs in one grid cell, it may flow downstream and cause flooding in adjacent cells. This routing effect becomes particularly important for fluvial flash floods in intermediate-sized catchments (100-500 km²), as seen, for example, during the severe flash floods occurred in Valencia in October 2024. Although ERA5's large grid cells ($\sim$31 km) may typically contain flash floods within their boundaries, downstream propagation can occur, with flash floods extending beyond the grid cell receiving rainfall and affecting neighbouring downstream grid-boxes.

6.2.2.3 Repeated nested cross-validation for model training

An important and long-standing concern in model training is *overfitting* (Ying, 2019). For predictive goals, overfitting degrades the generalisation of predictive performance to new data, and cross-validation is a technique that can help train models while limiting the risk of overfitting.

Simple cross-validation to estimate model's generalisation capabilities

Traditional model training works by splitting the available data into a set of *training* and *test* sets (Figure 6.2) where the model is fit to the training data and subsequently assessed based on its predictions to the test data (Hastie et al., 2009). By repeating this process for *k_outer* number of splits (Figure 6.2, *outer folds*), the average predictive performance of one or more models can be estimated. The splits are created with the function *RepeatedStratifiedKfold* in the SciKit-Learn Python. This function creates folds that preserve the original class distribution within each outer fold of the data partition[5], and repeats the partitioning n_repeats times.

Adoption of nested cross-validation to avoid data leakages when also tuning the model's hyperparameters

Cross-validation may also be used to estimate the model's hyperparameters. However, simple cross-validation uses the *same data* for model selection and hyperparameter tuning, which may introduce *data leakage*, thereby causing overfitting and compromising model generalisation capabilities (Sasse et al., 2025). The magnitude of this effect depends on the dataset size, the balance between the frequency of binary events in the dataset, and the model's stability (Sasse et al., 2025). A *nested* cross-validation framework provides unbiased estimates of model performance when using severely imbalanced datasets, whilst simultaneously optimising hyperparameters. Each of the *outer folds* is divided into an *outer test fold* and an outer training fold, which is split a *k_inner* number of times to create an *inner training fold* and an *inner validation folds*, preserving again the original class distribution within each inner fold (Figure 6.2). Hyperparameters are tuned over *inner training folds* and tested *inner validation folds*. When the hyperparameters are tuned, model generalisation is then tested over the *outer test folds*.

[5] In standard k-fold cross-validation, data points are randomly assigned to folds without consideration of their class membership, which can result in significant variations in class proportions across folds, particularly problematic for imbalanced datasets. Stratified k-fold cross-validation addresses this limitation by ensuring that each fold maintains approximately the same percentage of samples from each class as the complete dataset.

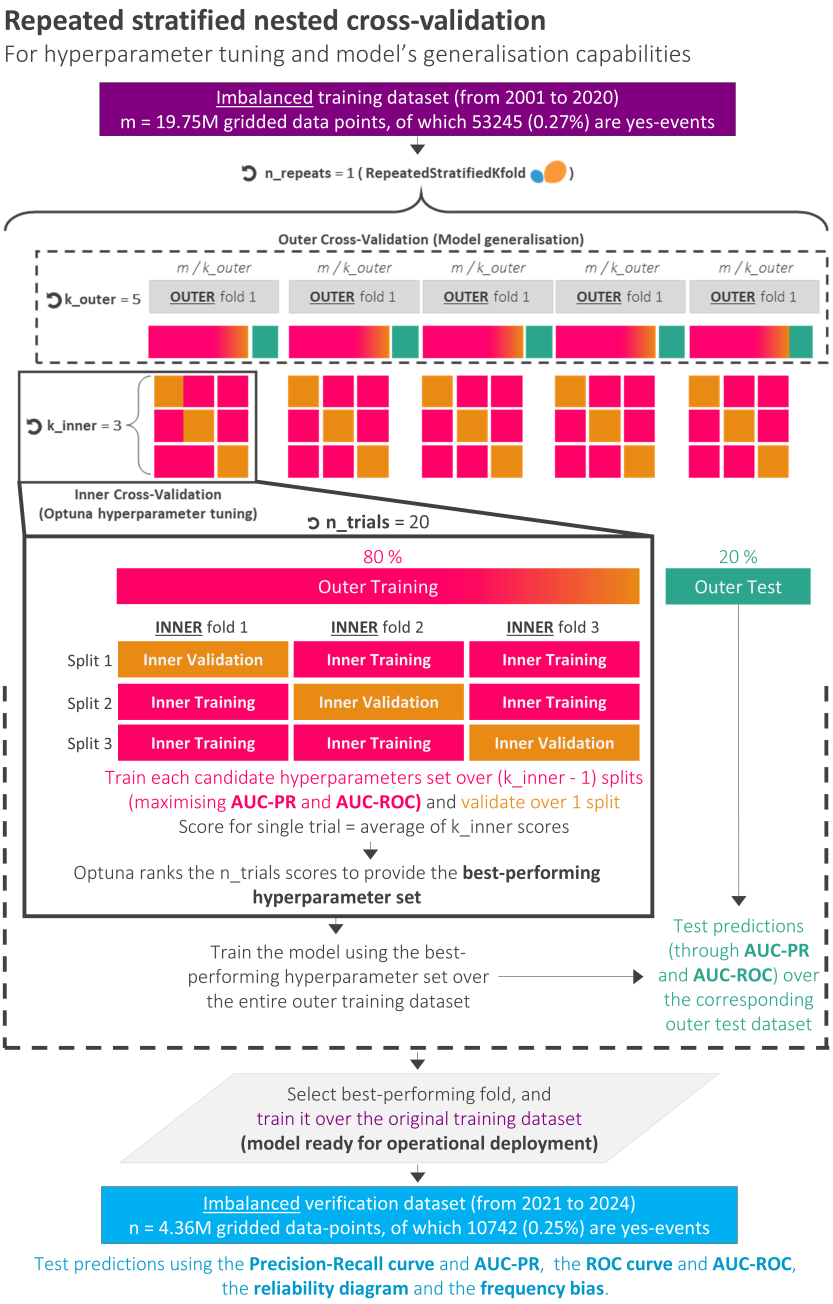

Figure 6.2: Workflow for the repeated nested cross-validation. The outer cross-validation loop utilises Scikit-Learn's "RepeatedStratifiedKFold" function to create k_outer = 5 outer folds (grey blocks) across n_repeats = 1 iterations. Each *outer fold* maintains the class distribution of the *training dataset*, and it is split into an outer training dataset (80%, blocks in shades of pink and orange) and an *outer test dataset* (20%). Within each outer fold, a Bayesian hyperparameter tuning is performed employing the Optuna library through an inner cross-validation procedure over n_trial = 20 repetitions. Each trial evaluates candidate hyperparameters by training on *inner training folds* and validating on *inner validation folds*, with performance measured as the mean AUC-ROC or AUC-PR. The optimal hyperparameter set, identified by maximising the selected evaluation metric, is used to train the final model on the complete outer training subset. Model performance is assessed on the held-out *outer test fold* using AUC-ROC and AUC-PR. The best-performing fold is retrained on the original *training dataset* for operational deployment. Independent, more extensive verification of the data-driven predictions is performed using the *verification dataset*, considering the Precision-Recall curve and AUC-PR, the ROC curve and AUC-ROC, reliability diagrams, and frequency bias.

6.2.2.4 Hyperparameter tuning

Hyperparameter tuning: Optuna implementation

The hyperparameter tuning was conducted using the Python library Optuna (Akiba et al., 2019). The framework implements a Bayesian optimisation algorithm, primarily utilising the Tree-structured Parzen Estimator (TPE) sampler, which models the relationship between hyperparameters and objective function values to navigate high-dimensional search spaces efficiently. This approach significantly outperforms traditional grid search and random search methods, particularly when computational resources are limited or when the hyperparameter space is complex and continuous. The framework's architecture enables dynamic construction of search spaces where hyperparameters can be conditionally dependent on one another. Optuna's pruning capabilities represent a crucial innovation for computationally intensive tasks, allowing early termination of unpromising trials based on intermediate performance metrics. Moreover, the MedianPruner implementations monitor trial progress and eliminate configurations that are statistically unlikely to surpass previously observed performance, thereby focusing computational resources on promising regions of the hyperparameter space. This feature proves particularly valuable when training deep neural networks or large ensemble models, where individual trial evaluation may require substantial time. Optuna's integration with popular machine learning frameworks such as XGBoost, LightGBM, CatBoost, and TensorFlow handle framework-specific optimisations such as early stopping criteria and validation monitoring, whilst maintaining compatibility with Optuna's pruning mechanisms. Finally, the comprehensive logging and visualisation capabilities facilitate post-hoc analysis of optimisation trajectories, parameter importance assessment, and convergence diagnostics, providing valuable insights into model behaviour and hyperparameter interactions.

The Bayesian optimisation process conducted through Optuna evaluated n_trial = 20 distinct hyperparameter configurations, with each trial exploring different regions of the search space guided by the Tree-structured Parzen Estimator algorithm.

6.2.2.5 Loss functions

The training framework implements a dual-strategy approach to address class imbalance through loss function configuration, enabling empirical determination of whether class weighting improves predictive performance for the specific characteristics of flash flood data.

Standard loss functions for balanced datasets

For balanced datasets or when class imbalance is not a primary concern, the standard *binary cross-entropy (BCE) loss function* is employed without modification. The standard BCE treats the binary classes equally, computing the negative log-likelihood of the predicted probabilities without any weighting mechanism. For tree-based models, the equivalent implementations include *binary:logistic* objective for XGBoost, *binary* objective for LightGBM, and *Logloss* objective for CatBoost. These standard formulations assume that misclassification costs are symmetric between classes and that the training data adequately represents the true class distribution, making them suitable when positive and negative instances occur with comparable frequency.

Specialised loss functions for imbalanced datasets

Due to the severe class imbalance inherent in the problem considered in this thesis, the *weighted binary cross-entropy (W-BCE) loss function* is also considered. It assigns differential importance to minority class instances through an optimisable positive class weight parameter (i.e., the class weights are themselves a hyperparameter optimised with Optuna), within the range [1.0, 10.0]. This weighting mechanism compensates for the scarcity of positive examples by increasing their contribution to the overall loss calculation, thereby preventing the model from converging to a trivial solution that simply predicts the majority class, i.e., it penalises misclassifications of rare flash flood events more heavily than false positives. The implementation manifests differently across frameworks: XGBoost and LightGBM utilise the *scale_pos_weight* parameter to directly multiply the loss contribution of positive instances, whilst CatBoost employs the *CrossEntropy loss function* with inherent class weighting capabilities. For neural networks, the Keras implementation applies *class weights during batch gradient computation*, effectively rebalancing the optimisation landscape to ensure adequate representation of rare events.

6.2.3 Objective verification framework

***Similarities* with the objective verification framework in Chapter 5**

The objective verification framework used in this study is similar to that in Section 5.2. Specifically, the reader is referred to Chapter 5 for a detailed description of how point flash flood impact reports are converted into gridded observational fields (Figure 5.3) and what scores are used to evaluate the reliability and discrimination ability of the data-driven predictions of areas at risk of flash floods (Figure 6.4).

***Differences* between the objective verification framework in Chapter 5**

Three key differences distinguish the objective verification framework

Hydro-meteorological data-driven predictions of areas at risk of flash floods
Defining yes- and non-events

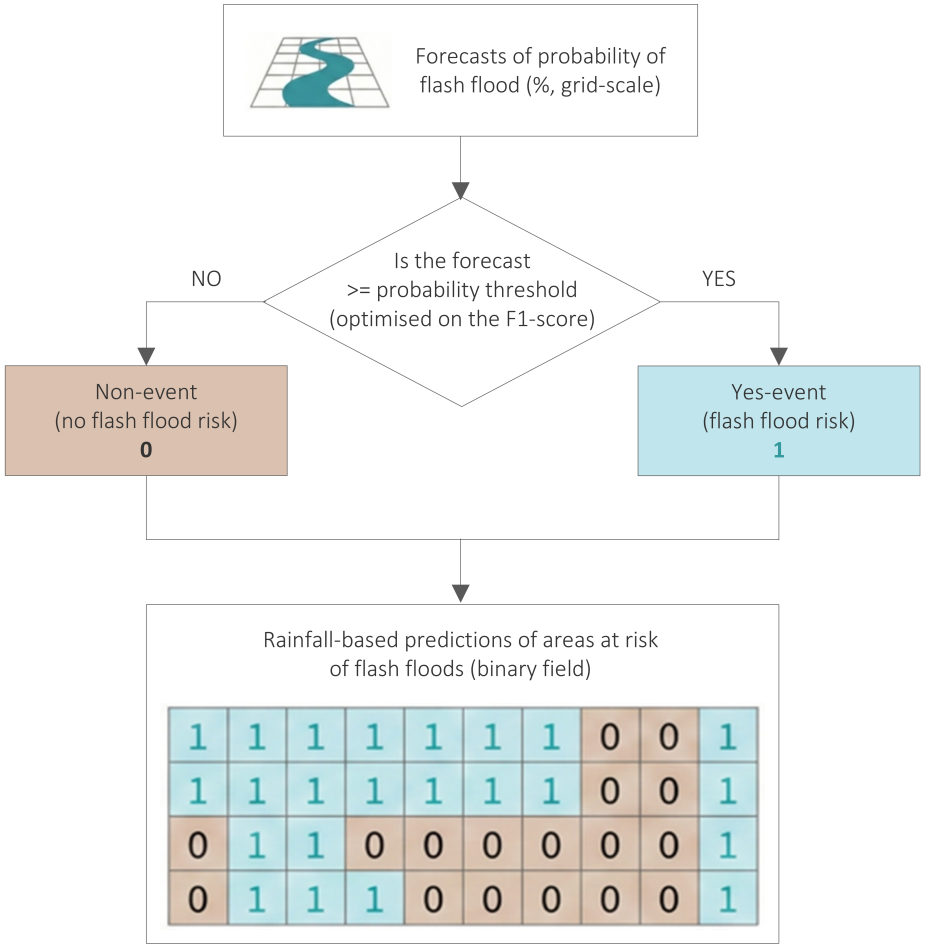

Figure 6.3: Schematic on how yes- and non-events are defined for the forecasts of probability of flash flood (in %, at grid-scale). The forecast of probability of flash flood (%) for each grid box is compared against a corresponding probability threshold (optimised on the F1-score). If the forecast exceeds or is equal to the threshold, it is classified as a "yes-event" (value 1, shown in light green), indicating a risk of flash flooding. Conversely, forecasts below the threshold are classified as "non-events" (value 0, shown in light brown). The resulting output is a binary field representing the predictions of areas at risk.

employed here from that presented in Chapter 5. First, for the data-driven predictions evaluated in this chapter, the verifying threshold is a probability value above which a grid-box is classified as a yes-event. Conventionally, a threshold of 50% is applied. However, for rare events such as flash floods, this would yield very few yes-event classifications, rendering verification insensitive to model performance. The threshold is therefore optimised on the F1-score, which balances precision and recall by penalising false alarms and missed events equally (Hancock et al., 2022). This approach identifies the decision boundary that maximises predictive skill given the

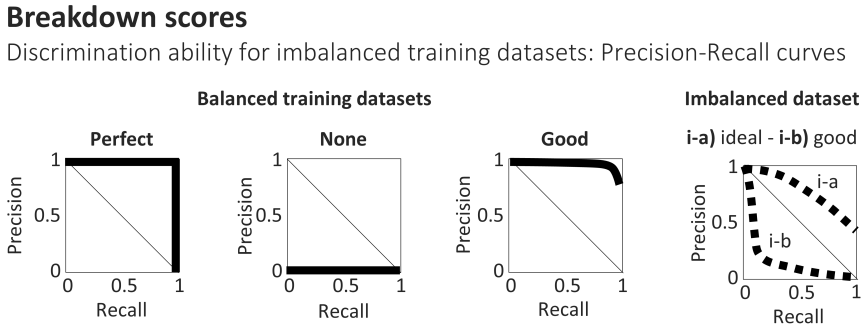

Figure 6.4: Breakdown score to assess discrimination ability for imbalanced training datasets: Precision-Recall curves. Examples of Precision-Recall curves for forecasts with perfect, none, and good discrimination ability (for balanced datasets). The figure also shows the typical precision-recall curves for imbalanced datasets, with ideal (i-a) and good (i-b) discrimination ability.

inherent trade-off between detection rate and false alarm rate. Second, the precision-recall (PR) curve is introduced to complement the discrimination analysis provided by ROC curves. PR curves are better suited than ROC curves for evaluating predictions trained on imbalanced datasets, as they focus on the minority class and are more sensitive to changes in false alarm rates when true positives are rare (Saito and Rehmsmeier, 2015; Juba and Le, 2019; Sofaer et al., 2019). The PR curve is built from the probabilistic contingency table in Figure 5.4, mapping precision (in the y-axis) against recall (in the x-axis), computed from equations 6.1 and 6.2:

$$\text{precision} = \frac{\text{H}}{\text{H} + \text{FA}} \quad [\text{values between } 0 \text{ and } 1] \tag{6.1}$$

$$\text{recall} = \frac{\text{H}}{\text{H} + \text{M}} \quad [\text{values between } 0 \text{ and } 1] \tag{6.2}$$

It is worth noting that recall is better known in hydro-meteorology as hit rate (Equation 5.3). Moreover, PR curves are more commonly known in meteorology as ”Performance Diagrams” and have been primarily applied to deterministic predictions (Taylor, 2001). Finally, 99% confidence intervals were computed via bootstrapping with replacement over 1000 repetitions for all scores. However, in this case, the confidence intervals were so small that they were not included in the figures presenting the verification results.

6.2.4 Physical interpretation of the data-driven model outputs: Shapley values

Shapley values, derived from cooperative game theory, have emerged

Shapley values enable interpretability by decomposing predictions into individual feature contributions

as a powerful technique for interpreting complex data-driven models by quantifying the contribution of each input feature to individual predictions (Rozemberczki et al., 2022). Shapley values decompose a model's prediction into additive contributions from each predictor variable, revealing how each model feature - in this case, rainfall, antecedent soil moisture, orography slope, and vegetation coverage - combines to produce the final prediction - i.e., areas at risk of flash floods. This decomposition provides crucial physical insights into the decision-making process of the data-driven model. In this way, the model becomes *interpretable* because it is possible to assess which variable yields the most importance in defining a prediction and how a model feature modulates the impact of another one. Moreover, it is possible to diagnose potential issues where the model may be relying on spurious correlations.

6.3 Results

6.3.1 Model training

Model's hyperparameter optimisation

The hyperparameter optimisation history plot (Figure 6.5) shows the overall good performance of the Optuna library in identifying quickly (within the first 10 trial, orange lines) and with relatively small variations (grey lines) the sets of hyperparameters that maximise the chosen evaluation metrics (AUC-ROC and AUC-PR). Performance remains fairly consistent between different outer folds (lines in shades of grey and orange). The optimisation histories for AUC-ROC (Figure 6.5a-l) and AUC-PR (Figure 6.5m-x) are very similar, with plots for loss functions specific for imbalanced datasets (Figure 6.5g-l and s-x) showing more variability, but in general better overall performance, than those for more general loss functions (Figure 6.5a-f and m-r).

Optuna's training times

The training times, defined as the total wall-clock time required for Optuna to complete hyperparameter optimisation for each outer fold, including all trial evaluations and final model training, show substantial computational differences across model architectures (Figure 6.6). Decision-tree-based implementations (except for CatBoost) demonstrate more efficient training times. On average, the training time per outer fold remains between 75 and 100 seconds, with peaks that do not exceed 500 seconds. CatBoost's training times are around 500 seconds per outer fold, with peaks reaching 2000 seconds. The feed-forward neural network exhibits the longest training times among all models, with an average training time of around 2000 sec-

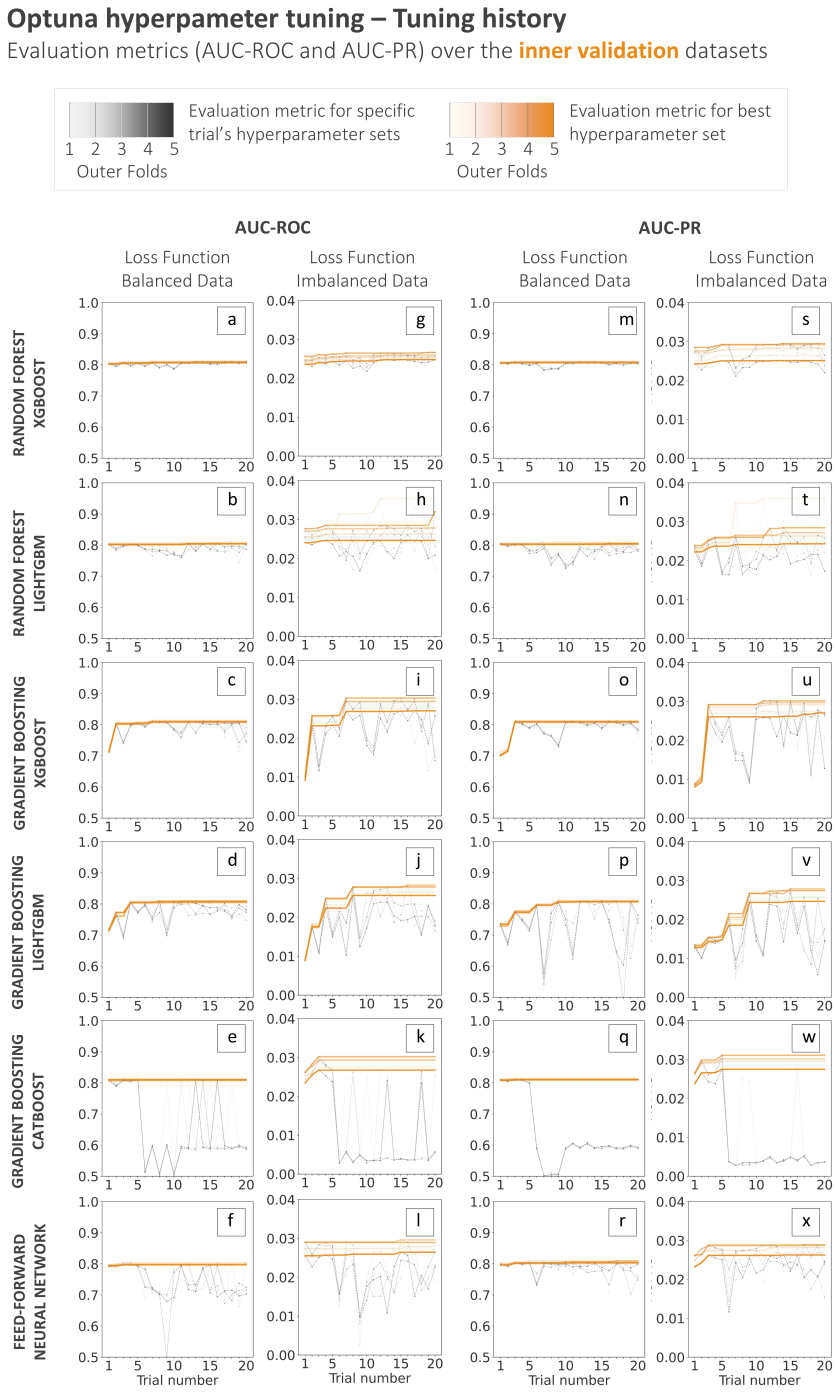

Figure 6.5: Optuna's hyperparameter optimisation history. Evolution of the two evaluation metrics (AUC-ROC, panels (a) to (l) - and AUC-PR, panels (m) to (x)) maximised during the 20 trials run over the *inner validation folds* to tune the hyperparameters of six data-driven models (from top to bottom): random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. The lines in shades of grey indicate individual trial performances, whilst lines in shades of orange highlight the best-performing hyperparameter set, identified by Optuna's Bayesian optimisation process. The shades of grey and orange represent the values of the evaluation metrics for each outer fold (lightest shade for the first outer fold and darkest for the latest). Panels (a) to (f) and (m) to (r) represents the results obtained using the standard binary cross-entropy loss functions - mostly used for balanced datasets - whilst panels (g) to (l) and (s) to (x) present the outcomes obtained with the weighted loss functions (specifically configured for imbalanced data).

onds and peaks ranging from 4000 to 6000 seconds. These times indicate that it requires ∼20 minutes per outer fold to optimise hyperparameters and train the decision-tree-based models (except for CatBoost), compared to ∼2.5 and ∼8 hours per outer fold, respectively, for CatBoost and the feed-forward neural network. The choice of loss function and evaluation metric shows minimal impact on training duration across all architectures. One notable characteristic in different panels of Figure 6.6 is the sudden increase in training time. No pattern for this behaviour is found, e.g., in outer fold n.1 for the gradient boosting XGBoost, trained with the AUC-ROC evaluation metric and with a loss function for a balanced dataset (Figure 6.6c) or in outer fold n.4 for the feed-forward neural network, trained with the AUC-PR evaluation metric and with a loss function for a imbalanced dataset (Figure 6.6x). Hence, this hieratic, random behaviour is likely to reflect differences in the composition of the training data for that fold (for instance, a higher absolute value of yes-events), which may require more trials for Optuna to converge or longer training times per trial.

Nested cross-validation: model generalisation

The close values for both evaluation metrics (AUC-ROC and AUC-PR) estimated over the *inner validation folds* and the *outer test folds* show the robustness of the nested cross-validation framework in mitigating overfitting during hyperparameter optimisation (Figure 6.7). The fact that outer test performance remains generally bounded or close to the performance ranges estimated over the inner validation folds demonstrates that Optuna's Bayesian optimisation successfully identified hyperparameter configurations with robust generalisation capabilities to previously unseen data. The comparative analysis reveals minimal divergence between models trained with standard binary cross-entropy and those employing weighted loss functions (formulated specifically for class-imbalanced datasets). This observation holds for both evaluation metrics. Across the evaluated models, performance metrics demonstrate remarkable consistency, with the notable exception of CatBoost. Specifically, mean AUC-ROC values cluster between 0.7 and 0.8, whilst CatBoost exhibits inferior performance with values between 0.6 and 0.7. Similarly, mean AUC-PR values range from 0.02 to 0.03 across most models, with CatBoost again demonstrating a lower performance with values between 0.01 and 0.02. For both metrics, random forest implementations exhibit the narrowest bands, followed by gradient boosting implementations (except for CatBoost, which displays the widest bands among all models) and the feed-forward neural network. The wider bands for CatBoost exhibit greater sensitivity to hyperparameter choices and require more careful tuning to achieve optimal results. Random

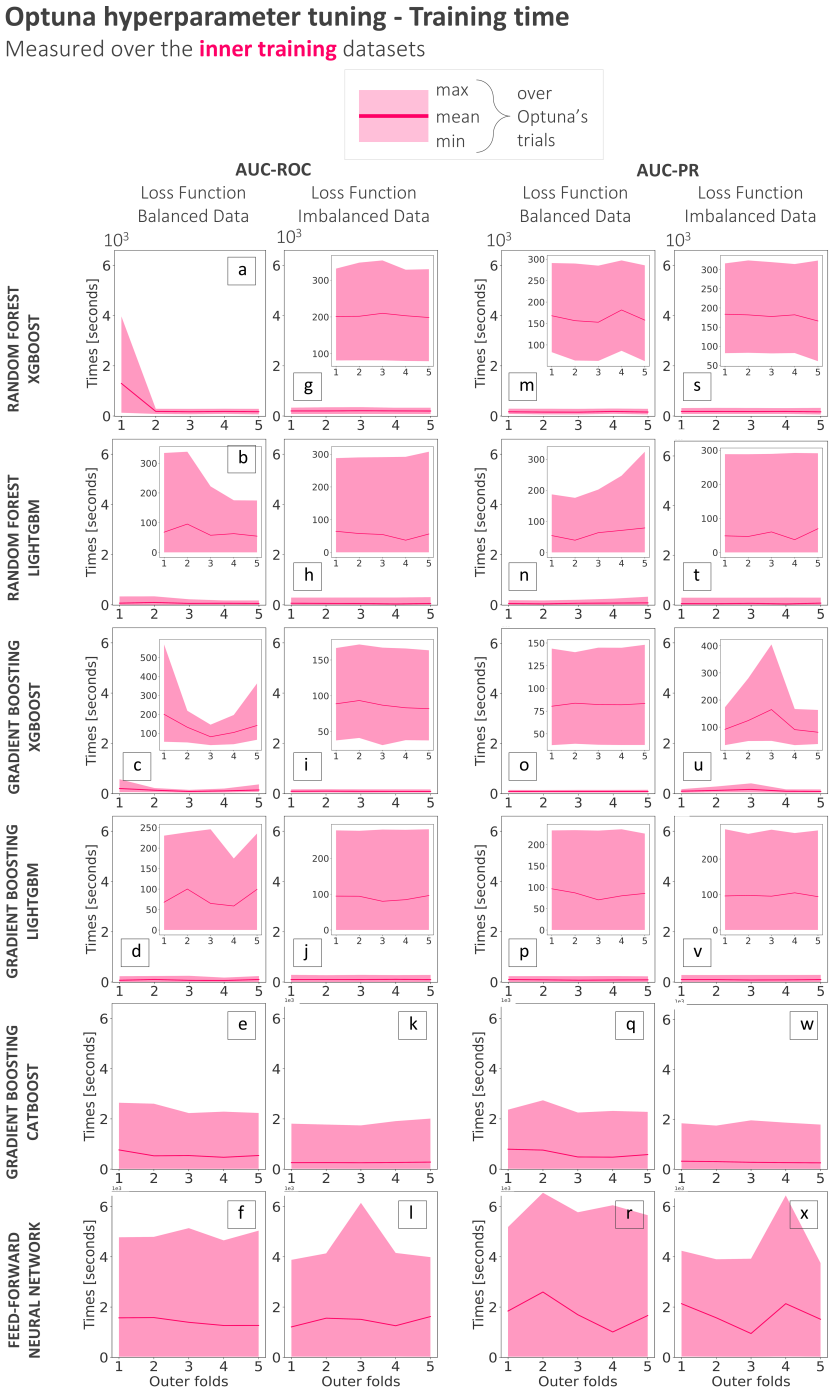

Figure 6.6: Optuna's training time. Evolution of training times (in seconds) for each k_outer=5 outer folds across the n_trials = 20 trials run over the *inner training folds* (the solid line represents the mean while the shaded area represents the minimum and maximum values). The training times for six data-driven models (from top to bottom) are shown: random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. Training times are shown for both evaluation metrics (AUC-ROC - panels (a) to (l) - and AUC-PR - panels (m) to (x)) and loss function configurations (balanced and imbalanced datasets). Inset plots provide magnified views where appropriate.

forest implementations exhibit the most constrained performance bands, indicating robust hyperparameter spaces that yield consistent results across outer folds. Gradient boosting implementations (except for CatBoost) and the feed-forward neural network demonstrate intermediate variability. CatBoost exhibits the broadest performance bands amongst all evaluated models, suggesting that CatBoost's hyperparameter space possesses heightened sensitivity, necessitating more meticulous optimisation procedures to achieve competitive performance levels. As observed for training times, certain outer folds exhibit anomalous behaviour, likely reflecting the same underlying variability in data composition across folds.

6.3.1.1 Hyperparameter importance

Hyperparameter importance: gradient boosting implementations

The hyperparameter importance analysis for XGBoost and LightGBM gradient boosting implementations reveals that maximum depth and learning rate consistently emerge as the most influential parameters. This aspect suggests that optimal performance depends fundamentally on striking a balance between model complexity and convergence dynamics. Maximum depth controls the model's capacity to capture complex hydro-meteorological interactions, whilst learning rate determines the magnitude of iterative corrections, requiring careful calibration to preserve gradient signals from rare positive events. LightGBM demonstrates additional sensitivity to the number of estimators due to its leaf-wise tree construction, producing more informative individual trees that necessitate precise ensemble size optimisation. In contrast, XGBoost's level-wise approach generates simpler trees that plateau predictably, reducing the criticality of this parameter. Notably, the positive class weight parameter exhibits a generally lower influence across both implementations, suggesting that structural parameters governing tree complexity and learning dynamics exert a greater impact on minority class detection than explicit re-weighting strategies. Thus, this emphasises the primacy of architectural optimisation over class balancing approaches. CatBoost exhibits considerable variability in the hyperparameters that most strongly influence model performance, with this variability dependent on both the chosen evaluation metric and the loss function configuration. When optimising for AUC-ROC, a set of parameters proves critical (e.g. depth and learning rate), yet these same parameters have minimal impact when considering AUC-PR. This divergence likely reflects the fundamentally different aspects of model performance that each metric captures: AUC-ROC rewards the classification of yes-events even if that leads to unreliable predictions (too many false alarms), whereas AUC-

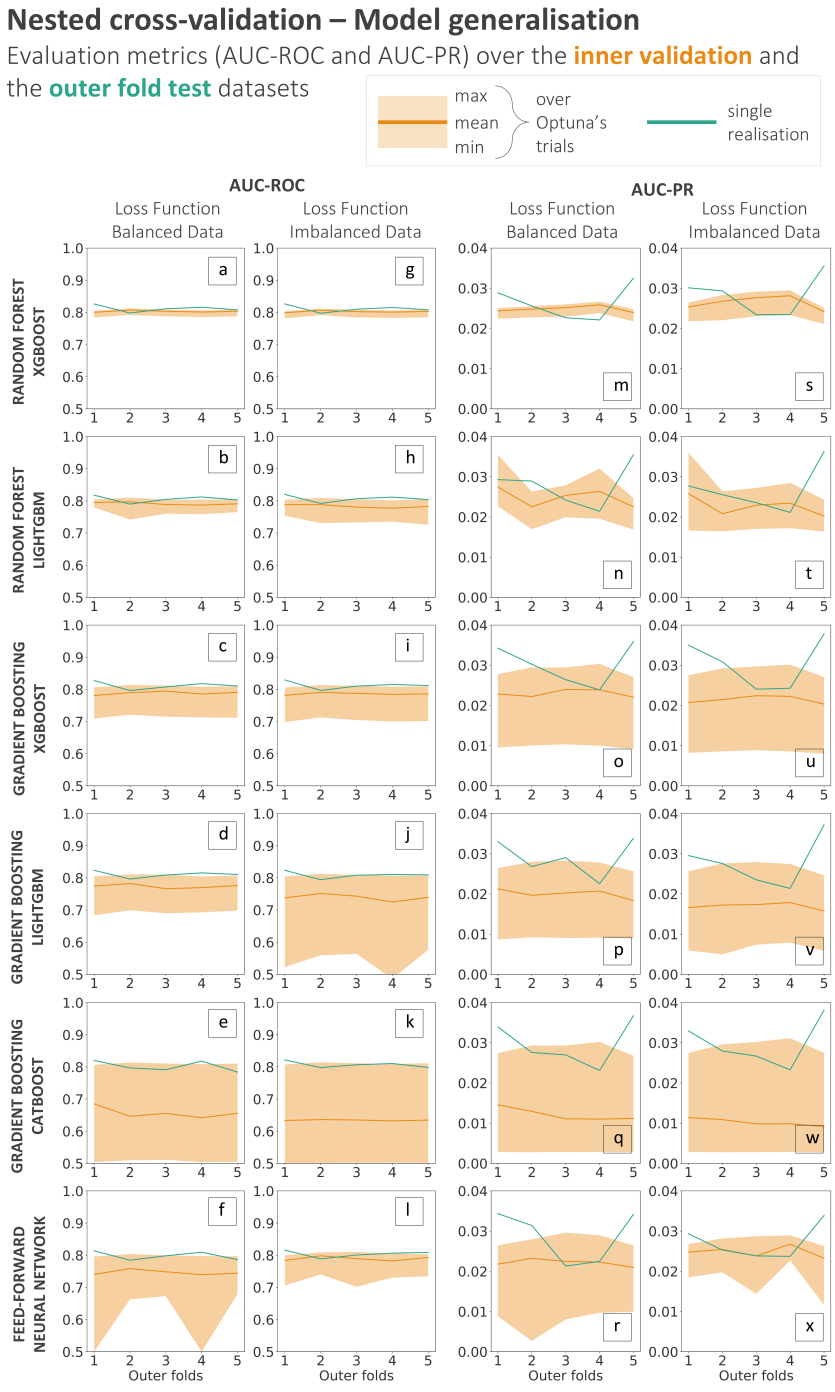

Figure 6.7: Model generalisation from nested cross-validation Values of the two evaluation metrics (AUC-ROC, panels (a) to (l) - and AUC-PR, panels (m) to (x)) across the 20 trials run over the *inner validation folds* (the solid line represents the mean while the shaded area represents the minimum and maximum values) and the *outer test fold* (the solid line correspond to the single realisation per outer fold). Panels (a) to (f) and (m) to (r) represent the results obtained using the standard binary cross-entropy loss functions - mostly used for balanced datasets - whilst panels (g) to (l) and (s) to (x) present the outcomes obtained with the weighted loss functions (specifically configured for imbalanced data).

PR penalises false alarms more severely as it aims to keep predictions reliable. Furthermore, the implementation of weighted loss functions fundamentally alters the previously seen importance rankings, creating distinct optimisation priorities for balanced versus imbalanced scenarios. Unlike XGBoost and LightGBM, where maximum depth and learning rate consistently dominate, CatBoost lacks such universal governing parameters. The absence of consistent parameter hierarchies suggests that CatBoost's distinctive algorithmic approach generates a more complex hyperparameter space.

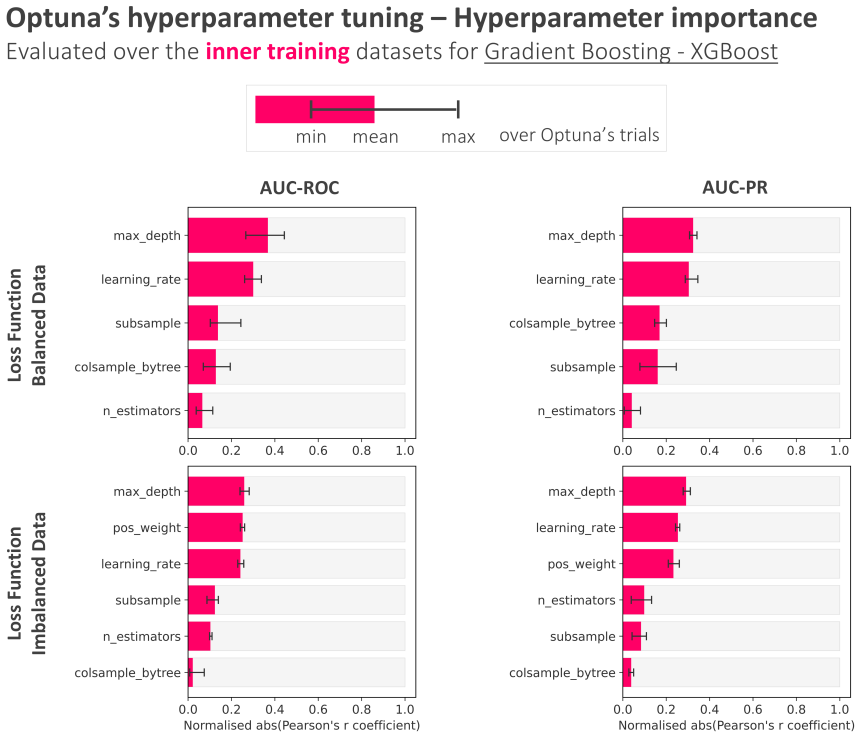

Figure 6.8: Optuna's hyperparameter importance for the XGBoost implementation of gradient boosting. Normalised absolute Pearson's correlation coefficients obtained for the n_trials = 20 trials run over the *inner training folds* (bars represent mean values, whilst error bars show the minimum and maximum values across the Optuna trials). Feature importance is shown for models trained with loss functions for balanced datasets and specific for imbalanced datasets, and for both evaluation metrics (AUC-ROC and AUC-PR).

Hyperparameter importance: random forest implementations

In contrast to gradient boosting implementations, random forest models exhibit inconsistent hyperparameter importance rankings across different loss functions and evaluation metrics, with parameters such as maximum depth, feature sampling ratios, and the number of estimators alternating in their relative influence depending on whether AUC-ROC or AUC-PR is optimised. This variability suggests that random forest hyperparameter spaces

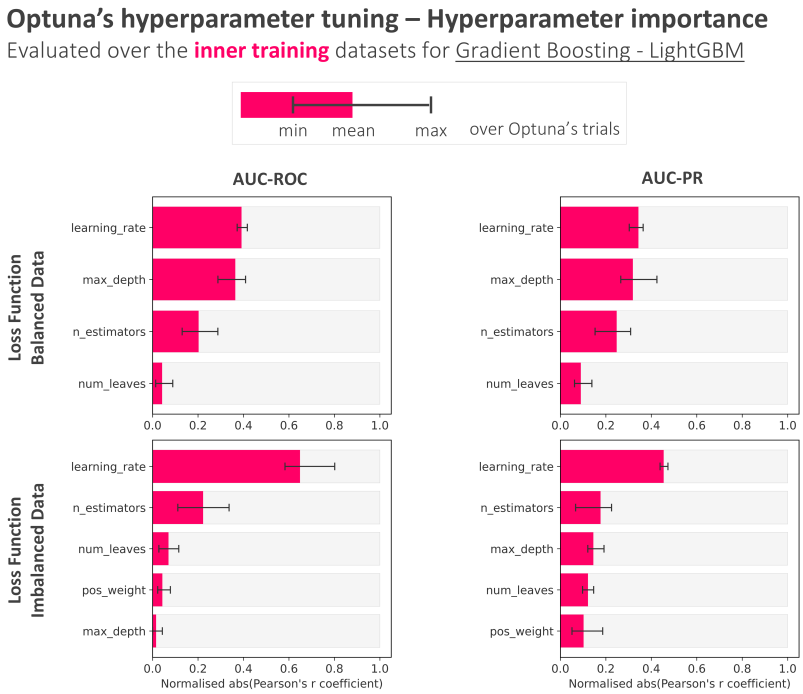

Figure 6.9: Optuna's hyperparameter importance for the LightGBM implementation of gradient boosting. Similar to Figure 6.8

.

possess multiple viable configurations that achieve similar performance through different mechanisms (some configurations may excel through deeper individual trees, whilst others compensate with more aggressive feature sampling or a larger ensemble size), making the optimisation landscape more flexible but potentially more challenging to navigate systematically.

Hyperparameter importance: feed-forward neural networks

The dominance of units_0, dropout_0, and learning rate for neural networks suggests that performance is primarily determined by the configuration of the first hidden layer, rather than the overall network depth. The parameter units_0 controls the initial representational capacity, determining how effectively raw hydro-meteorological features are transformed into meaningful intermediate representations for detecting rare flash flood patterns. This layer must balance sufficient complexity to capture non-linear relationships whilst avoiding overfitting to the sparse positive examples. The high importance of dropout_0 demonstrates that regularisation at this critical layer is essential for generalisation under extreme class imbalance. By randomly deactivating neurons during training, dropout forces the development of robust, redundant representations that prevent the model from memorising noise in the limited positive examples. The prominence of the

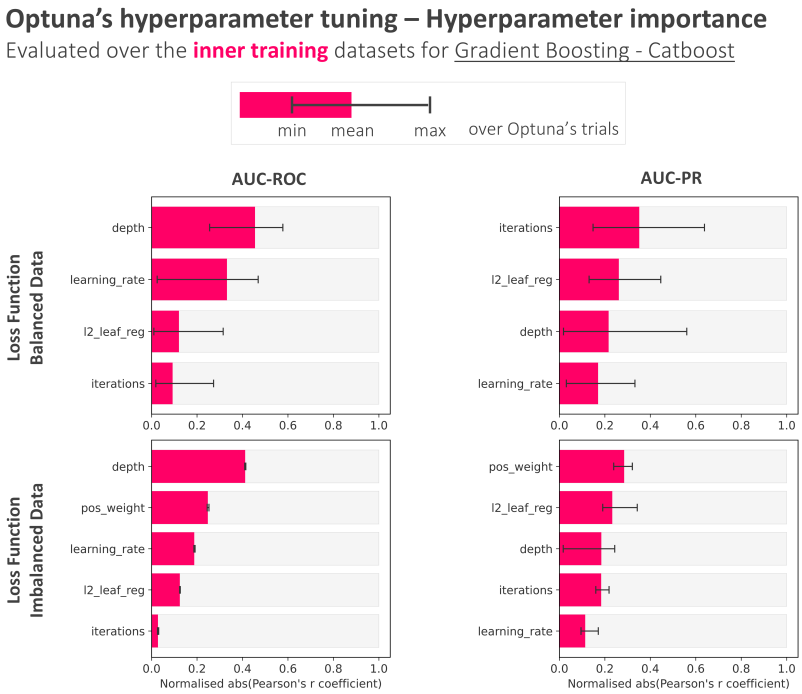

Figure 6.10: Optuna's hyperparameter importance for the CatBoost implementation of gradient boosting. Similar to Figure 6.8

.

learning rate parameter reflects the challenge of navigating an imbalanced dataset, where the gradient signal from rare positive examples can be easily overwhelmed by the abundance of negative cases. The diminished importance of deeper layer parameters suggests that shallow, well-regularised architectures may be the best choice for this application.

6.3.2 Verification results over reanalysis data

To assess the generalisation capabilities of the trained data-driven models, this section compares various performance metrics (presented in Section 6.2.3) between the training dataset, with data from 2001 to 2020, and the independent verification dataset, with data from 2021 to 2024.

Verification results over reanalysis data: overall scores (AUC-ROC, AUC-PR, and FB)

All data-driven models exhibit relatively stable AUC-ROC values around 0.8 for the verification dataset with minimal decrease from the AUC-ROC values obtained for the training dataset (Figure 6.14, first column). This behaviour is constant across both evaluation metrics (AUC-ROC and AUC-PR) and both types of loss function (general for balanced datasets and specific for imbalanced datasets). Hence, all data-driven models show

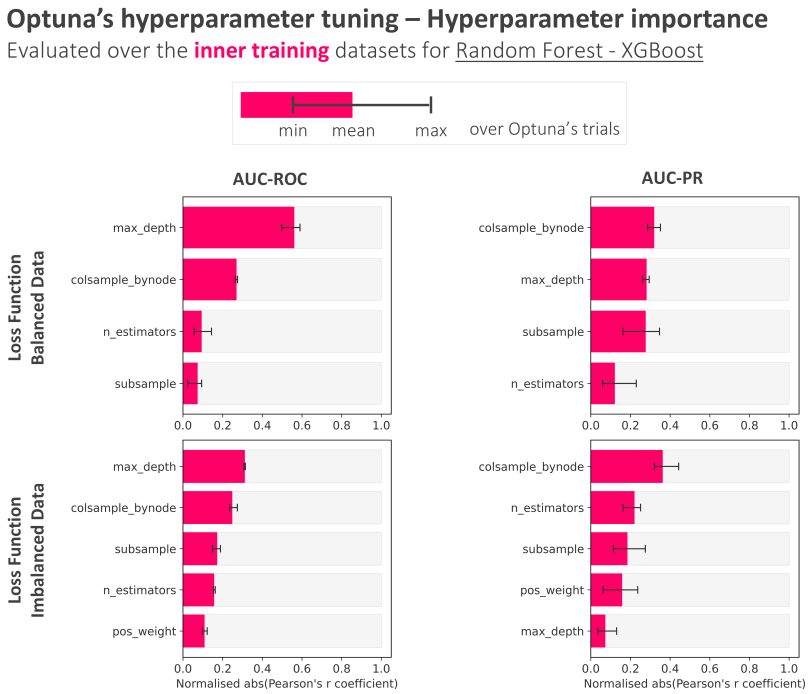

Figure 6.11: Optuna's hyperparameter importance for the XGBoost implementation of random forest. Similar to Figure 6.8

.

an overall discrimination ability that generalises well from training to verification datasets. The AUC-PR values demonstrate more variability across data-driven models (Figure 6.14, second column), with CatBoost showing the highest estimates overall, especially when hyperparameters are tuned maximising the AUC-ROC evaluation metric (for both types of loss function). For all models, however, values of AUC-PR remain relatively small (Close to 0). Even though all models show a tendency to slightly overestimate the frequency of areas at risk of flash flood (FB just above 1, Figure 6.14, third column), the FB values also show high variability across the data-driven models. When considering the generic loss function, the XGBoost and LightGBM implementations for random forest and gradient boosting obtain the closest values to 1, but increase of 20% when considering the loss function specific for imbalanced datasets, showing even greater FB than the other tested models (which maintain similar values than those obtained when considering the generic loss function).

Verification results over reanalysis data: breakdown scores (ROC curves)

Overall , ROC curves exhibit minimal degradation between training datasets and verification datasets (Figure 6.15). The ROC curves, however, reveal distinct patterns between models trained with balanced 6.15a-

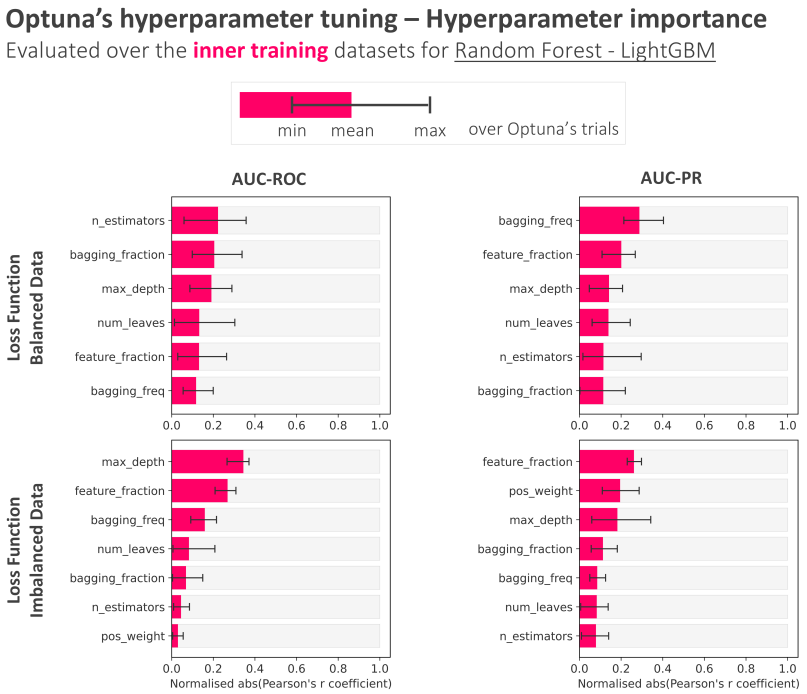

Figure 6.12: Optuna's hyperparameter importance for the LightGBM implementation of random forest. Similar to Figure 6.8

.

f and m-r) versus imbalanced loss functions 6.15g-l and s-x). The ROC curves computed for models employing balanced loss functions maintain a remarkably consistent shape across all data-driven models and evaluation metrics, as well as a consistent relative difference between ROC curves computed using the 1% discretisation (solid lines) and the 0.01% (dashed lines). The former ROC curves yield lower AUC-ROC values (AUC-ROC $\sim$0.7) than the latter (AUC-ROC $\sim$0.8) due to the "truncation effect" caused by stopping the ROC curve at the 1% probability threshold. The same does not hold when considering models trained with weighted loss functions. These ROC curves achieve higher hit rates, e.g., $\sim$0.8 for the XGBoost and LightGBM implementations of gradient boosting (Figures 6.15i-j) and $\sim$0.76 for the feed-forward neural network (Figures 6.15l) compared to $\sim$0.6 for their balanced counterparts (Figures 6.15c-d and f), but at the cost of much higher false alarm rates, e.g., >=0.5 for XGBoost and LightGBM and $\sim$0.2 for the feed-forward neural network compared to 0.025 in their balanced counterparts. Whilst XGBoost and LightGBM implementations of gradient boosting and the feed-forward neural network show consistent patterns across evaluation metrics (Figures 6.15g-h compared to Figures 6.15s-t), the random forest implementations display metric-dependent responses.

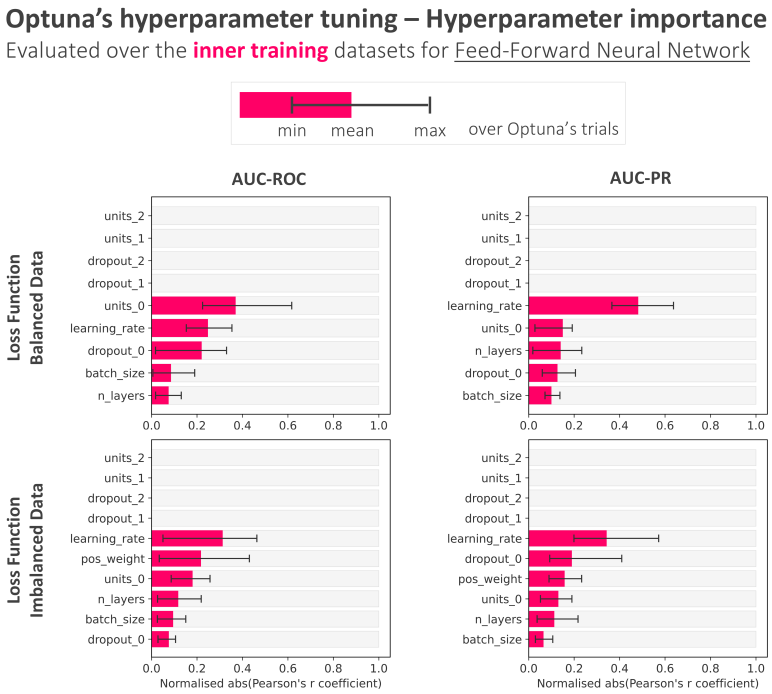

Figure 6.13: Optuna's hyperparameter importance for feed-forward neural network. Similar to Figure 6.8

.

Specifically, LightGBM Random Forest shows an altered behaviour when optimised for AUC-ROC, whereas XGBoost Random Forest responds differently under AUC-PR.

Verification results over reanalysis data: breakdown scores (Precision-Recall curves)

Overall , the precision-recall curves exhibit minimal degradation between training datasets (in solid lines) and verification datasets (in dashed lines), with the major differences concentrated mainly over recall values smaller than 0.2 (Figure 6.16). All precision-recall curves, while obtaining fairly small values of AUC-PR (see Figure 6.14, second column), remain away from their corresponding lines of no skill (grey solid lines for the training dataset and grey dashed line for the test dataset, which mostly overlap). Whilst there are some differences between the precision-recall curves attributable to the differences in model training, this time (as opposed to what was seen for the ROC curves), the most considerable differences are observed between the models themselves. In all implementations of gradient boosting (Figures 6.16c-e, i-k, o-q, and u-v), the training datasets achieve high precision (between 0.6 and 1) over very small values of recall, while the verification datasets achieve a precision value between 0.2 and 0.6. Only CatBoost, trained with the weighted loss function, maintains

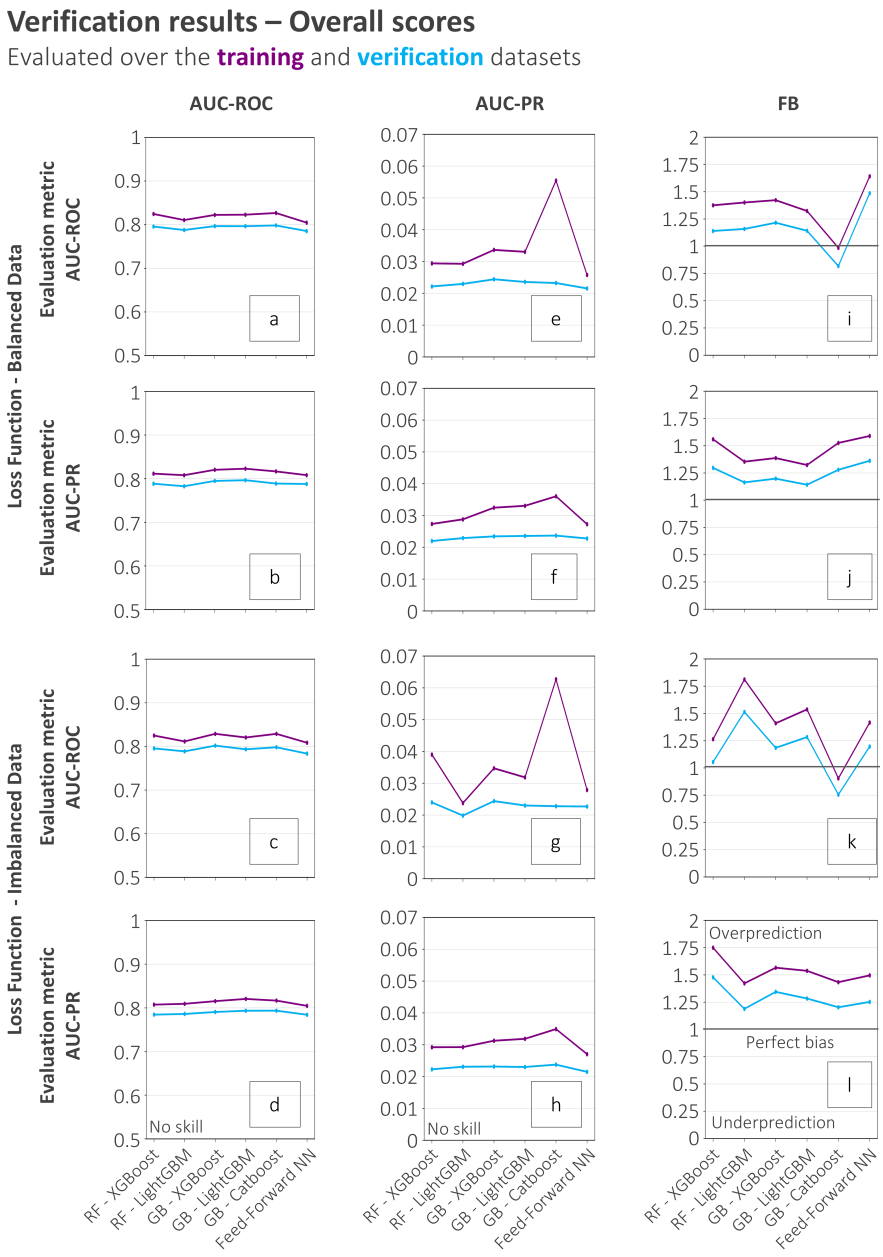

Figure 6.14: Verification results: overall scores The first, second, and third columns show, respectively, the estimates for the area under the ROC curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the frequency bias (FB) for the six considered data-driven models. The estimates of the overall verification scores are shown for the training dataset and the verification dataset. Results are presented for the models trained considering both types of loss functions and evaluation metrics.

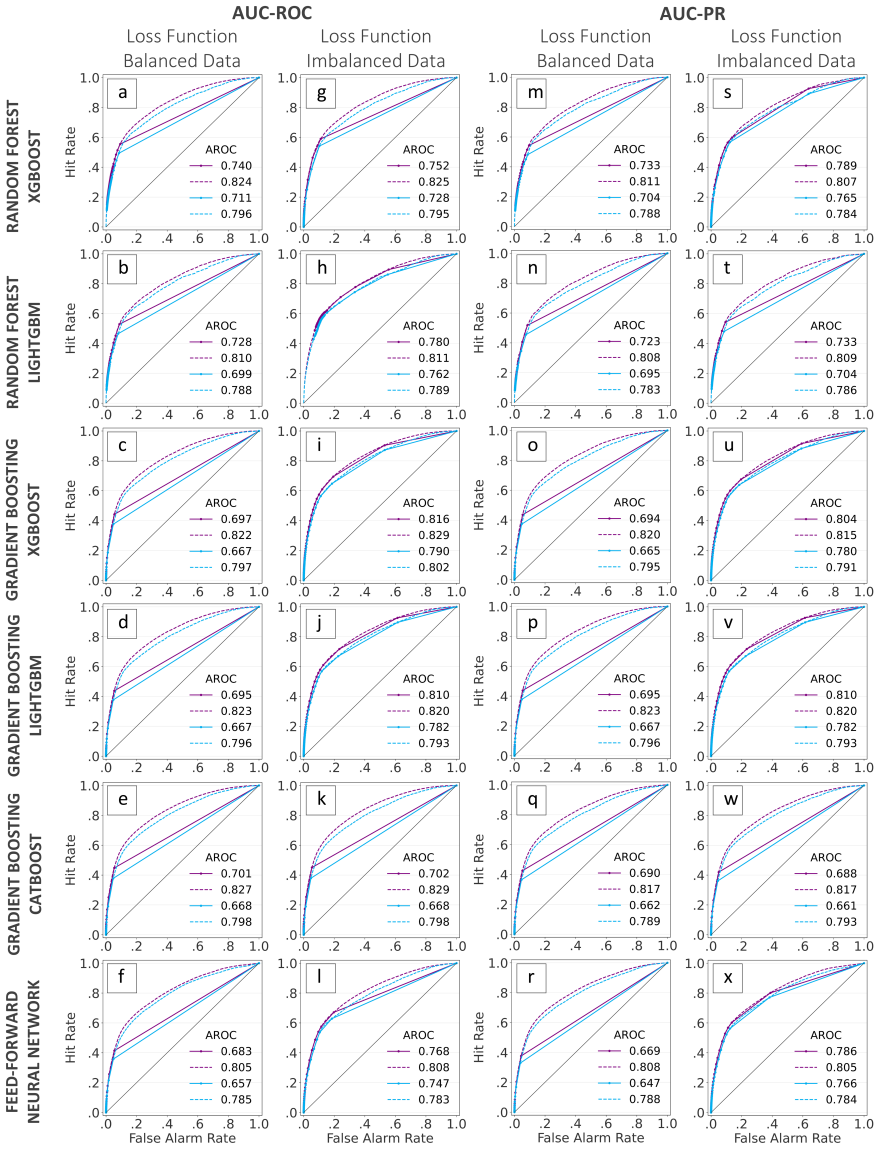

Figure 6.15: Verification results: breakdown scores (ROC curves) ROC curves computed for the training dataset and the verification dataset, for six data-driven models (from top to bottom): random forest XGBoost, random forest LightGBM, gradient boosting XGBoost, gradient boosting LightGBM, gradient boosting CatBoost, and feed-forward neural network. The solid lines represent ROC curves computed considering forecast probabilities discretised every 1%, whilst the dashed lines represent ROC curves computed using a finer discretisation of probabilities (0.01%). ROC curves are shown for the two evaluation metrics (AUC-ROC - panels (a) to (l) - and AUC-PR - panels (m) to (x)) and types of loss function configurations (balanced and imbalanced datasets) considered during the training of the data-driven models.

the value of 1. In the feed-forward neural network (6.16f, l, r, and x), the precision-recall curve remains virtually identical between the training datasets and the verification datasets, including at very low values of recall. Finally, both XGBoost and LightGBM implementations of random forest show very different precision-recall curves. LightGBM (6.16b, h, n, and t) shows a precision-recall curve that begins at recall values between 0.2 and 0.4, and precision values that do not exceed 0.1. XGBoost shows a similar behaviour when the model is trained using the generic loss function (6.16a and m), whilst the curves for the models trained with the weighted loss function (6.16g and s) shows a shape that is similar to that for the corresponding XGBoost implementation of gradient boosting (6.16i and u).

Verification results over reanalysis data: breakdown scores (Reliability diagrams)

Overall , the reliability diagrams exhibit minimal degradation between training datasets and verification datasets, but it increases with increasing forecast probabilities (Figure 6.17). Models trained with balanced loss functions (Figures 6.17a-f and m-r) demonstrate different calibration characteristics compared to those using weighted loss functions 6.17g-l and s-x). For models trained with balanced loss functions, the gradient boosting implementations and the feed-forward neural network yield reliable forecast probabilities under 10% (20% for XGBoost). At higher probability ranges, where predictions become less frequent and diagrams noisier, distinct patterns emerge. LightGBM and XGBoost show mostly reliable probabilities across all probability ranges (Figures 6.17c-d and o-p). CatBoost systematically underestimates forecast probabilities (Figures 6.17e and q). The feed-forward neural network shows contrasting behaviour depending on the evaluation metric used during hyperparameter tuning: overestimation occurs when tuned for AUC-ROC (Figure 6.17f), while forecast probabilities remain reliable when tuned considering AUC-PR (Figure 6.17r). Models trained with weighted loss functions (Figures 6.17g-l and s-x) display systematic overestimation across the entire probability spectrum, except for CatBoost, which maintains similar calibration patterns to its balanced function counterpart. Random forest implementations show systematic overestimations independent of the considered loss function 6.17a-b, g-h, m-n, and s-t).

6.3.3 Verification results over forecast data

Verification results over forecast data: overall scores (AUC-ROC, AUC-PR, and FB)

The area under the ROC curve (AUC-ROC, Figure 6.18a) and the area under the precision-recall curve (AUC-PR, Figure 6.18b) show a very close performance at day 1 (t+24) to that achieved over the reanalysis data (refer to Figure 6.14), and it gradually diminishes with increasing lead times. The

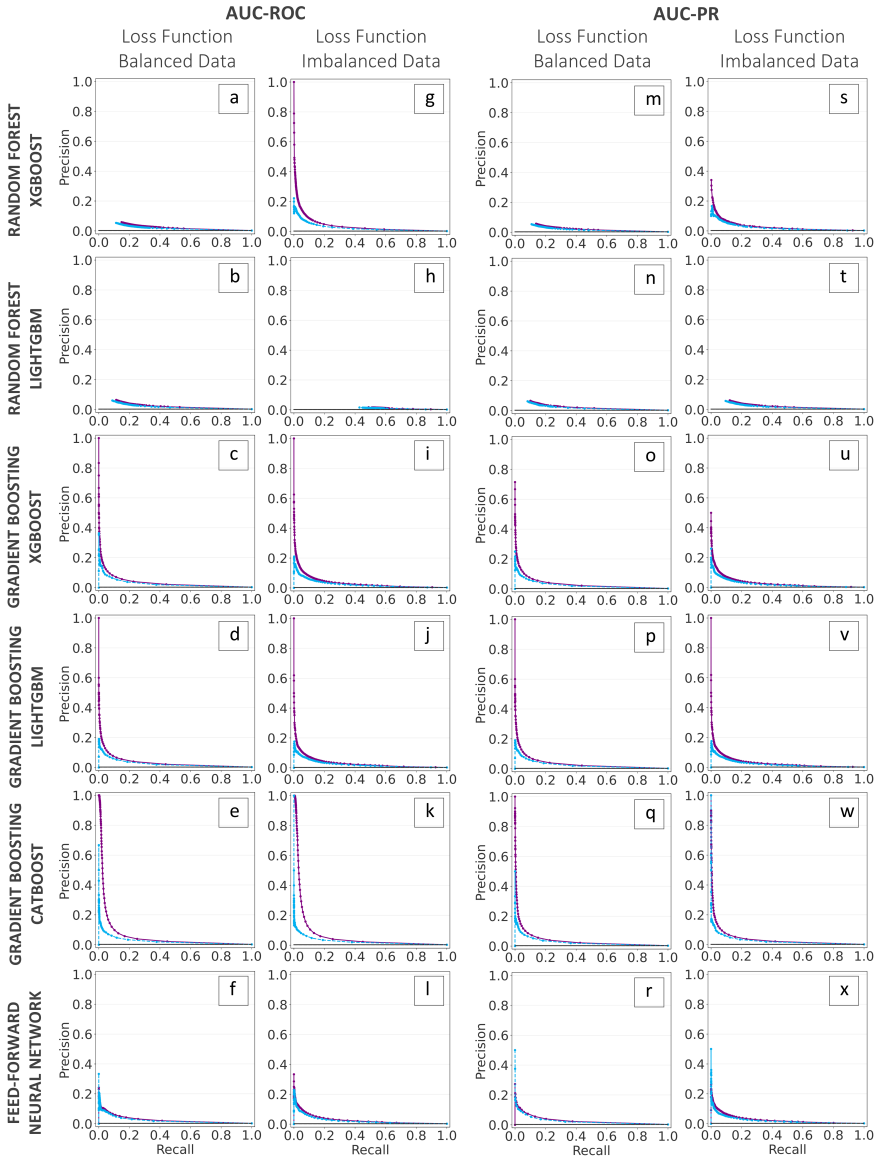

Figure 6.16: Verification results: breakdown scores (Precision-Recall curves) Similar to Figure 6.15 but for the Precision-Recall curve

frequency bias (FB) also slightly deteriorated with increasing lead times, but at day one (t+24), it shows a value that is twice the one obtained for the reanalysis data.

Verification results over forecast data: breakdown scores (ROC curves, precision-recall curves, and reliability diagrams)

The ROC curve for day 1 forecasts (t+24, Figure 6.18d) also shows a very similar behaviour to that for reanalysis data (refer to Figure 6.15),

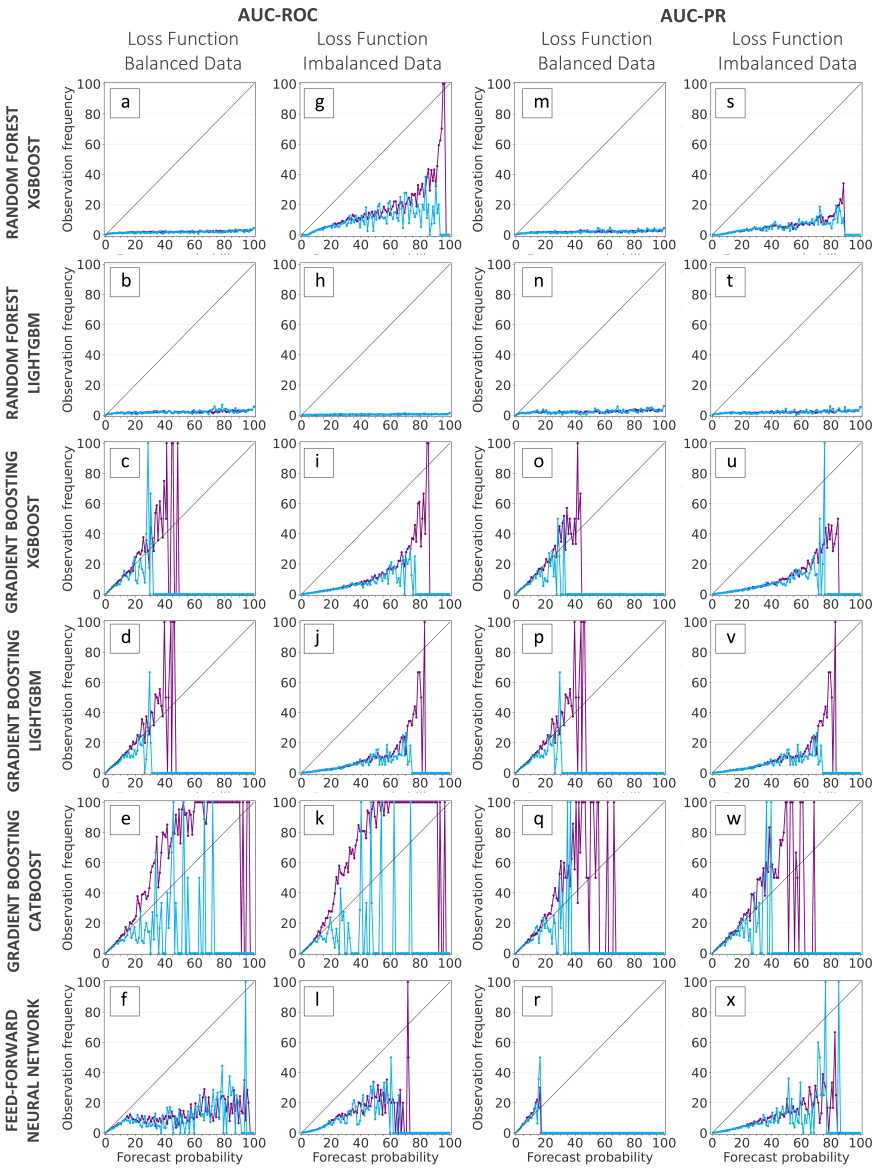

Figure 6.17: Verification results: breakdown scores (Reliability diagrams) Similar to Figure 6.15 but for reliability diagrams.

and shows a fairly small deterioration over increasing lead times (Figure 6.18e-f). The precision-recall curve for day 1 forecasts (Figure 6.18g) also shows a very similar behaviour to that for reanalysis data (refer to Figure 6.16), except for very small values of recall, where in the precision-recall curve built for the reanalysis data the precision is double. As in the ROC curve, the precision-recall curves also show only a fairly small deterioration

over increasing lead times (Figure 6.18h-i). As in the two previous scores, also the reliability diagram for day 1 forecasts (Figure 6.18j) shows a similar behaviour to that computed for reanalysis data (refer to Figure 6.17), with reliable probabilities for forecast probabilities under 10%. For increasing lead times, such a threshold reduces to $\sim$2%. For greater probabilities, the reliability diagrams tend to overestimate the observed frequencies of areas at risk of flash floods.

6.4 Physical interpretation of the data-driven model behaviour

For the sake of brevity and clarity, the following section will present SHAP-related plots exclusively for the XGBoost implementation of gradient boosting, trained with the balanced loss function and optimised for AUC-ROC. This representative configuration was selected because the feature importance patterns and SHAP value distributions are fairly consistent across all models, loss functions, and evaluation metrics examined in this study.

Physical interpretation of model behaviour with SHAP: global feature importance ranking

The global feature importance ranking (Figure 6.19a) shows that the rainfall variable for the probability of exceeding the 1-year return period emerges as the most important feature to identify areas at risk of flash flood, contributing to the 80% of the mean absolute SHAP values. Features regarding the vegetation cover (LAI), the orography steepness (sdfor), the antecedent soil moisture (swvl), and the rainfall probabilities of exceeding the 1-year return period in adjacent grid-boxes are also considered important by the model, contributing to 10 to 35% of the mean absolute SHAP values. The features related to the rainfall probabilities of exceeding the 50-year return period (*tp_prob_50* and *tp_prob_50_adj_gb*) are considered, overall, least important by the model to identify areas at risk of flash flood.

Physical interpretation of model behaviour with SHAP: dependency plots

The dependency plots show critical threshold behaviours in the model's decision-making process. The rainfall-related feature *tp_prob_1* (Figure 6.19b-d) contributes positively to the probabilities of having a flash flood in a grid-box. This contribution increases rapidly (from 0 to +3%) from probabilities between 0% and $\sim$40%. For greater probability values (>40%), the contribution of the rainfall parameters plateaus. For the rainfall-related feature *tp_prob_50* (Figure 6.19e-g), the behaviour is similar, but the threshold at which the contributions to the SHAP values plateau reduces to $\sim$7%. The

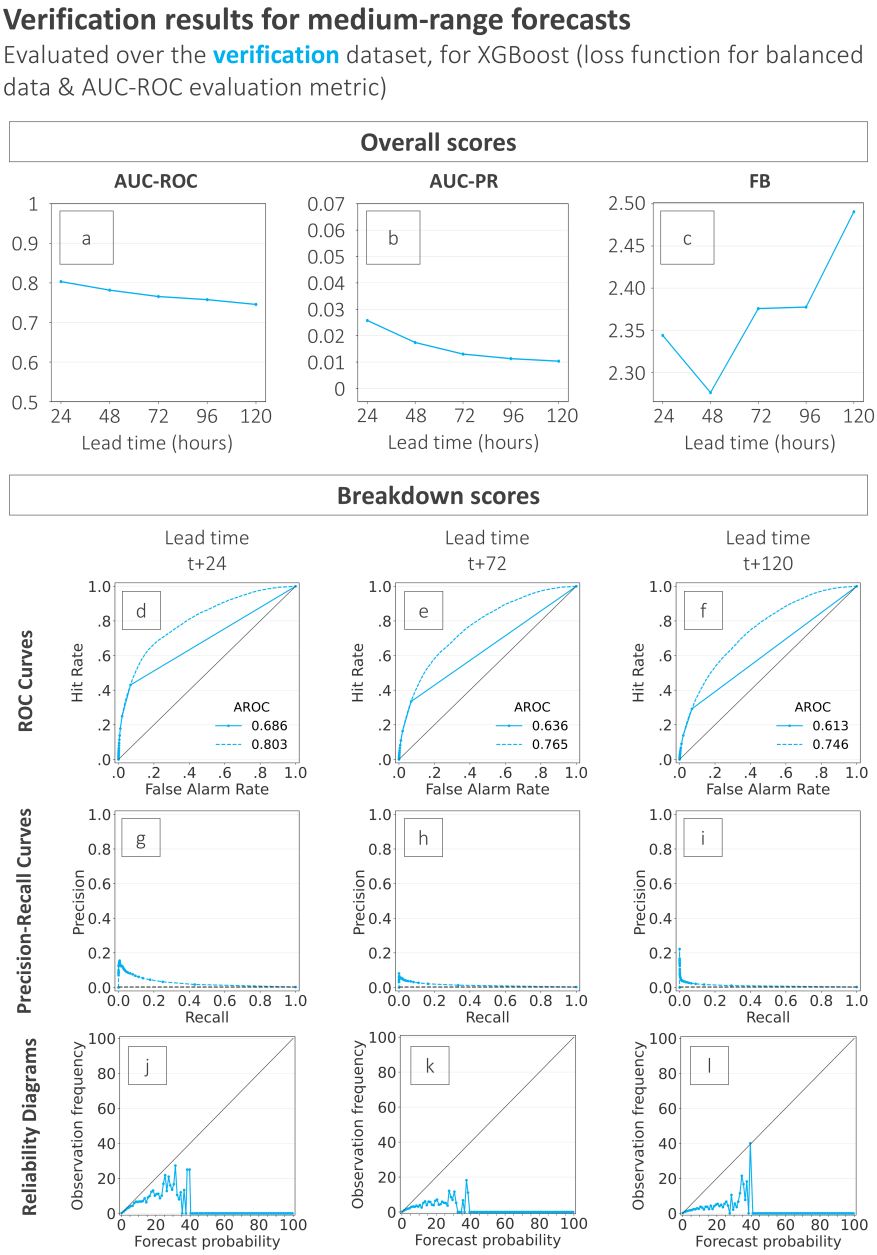

Figure 6.18: Verification results for medium-range forecasts for the XGBoost implementation of gradient boosting (trained with the loss function for balanced data and hyperparameters maximised for AUC-ROC). Panels (a) to (c) show the overall scores, respectively, for the area under the ROC curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the frequency bias (FB). The remaining panels show the breakdown scores, respectively, for the ROC curves (Panels (d) to (f)), the Precision-Recall curves (Panels (g) to (i)), and for the Reliability Diagram (Panels (j) to (l)).

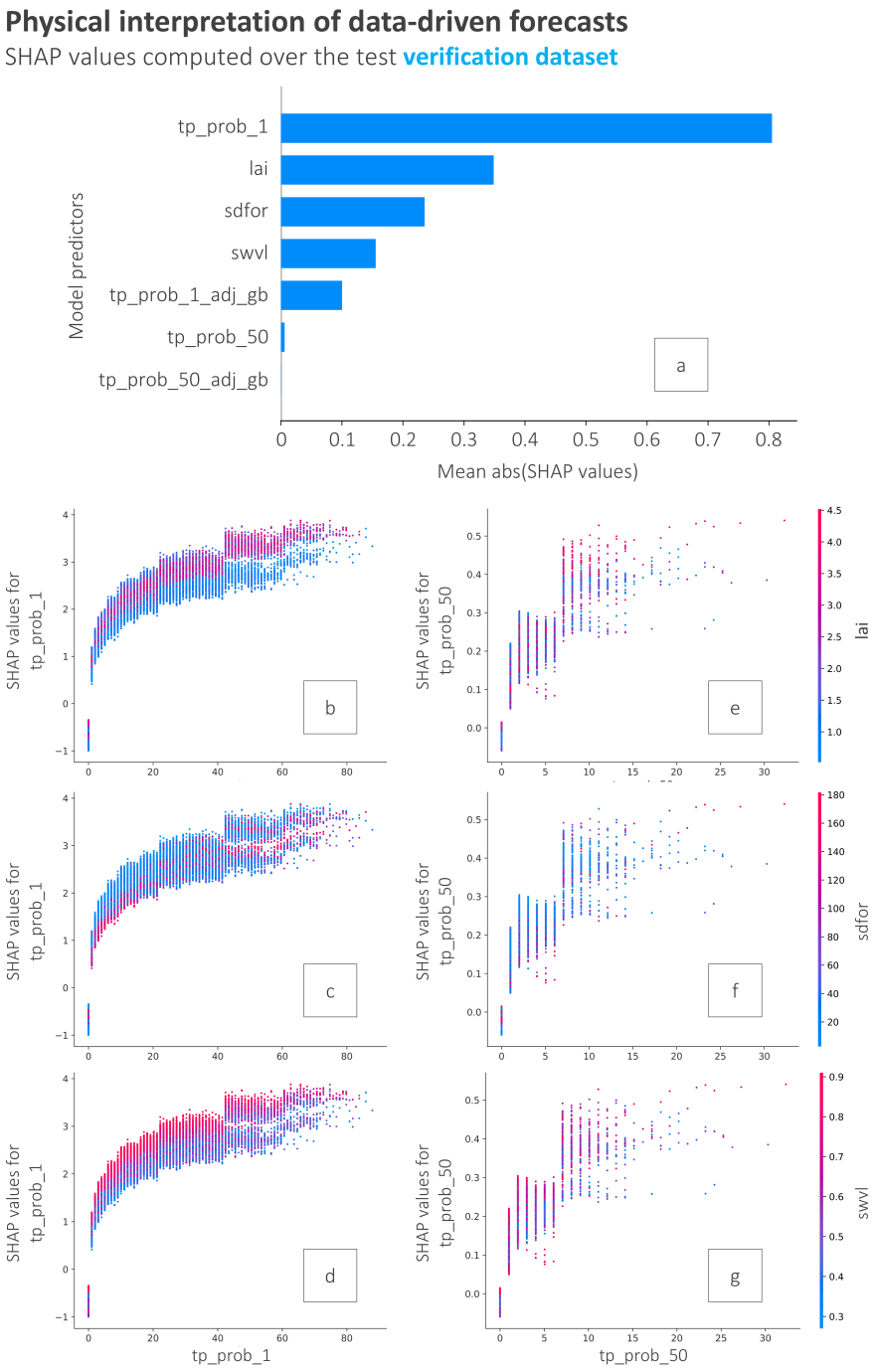

Figure 6.19: SHAP ((SHapley Additive exPlanations)) values over the verification dataset for the XGBoost implementation of gradient boosting (trained with the loss function for balanced data and hyperparameters maximised for AUC-ROC). Panel (a) shows the global feature importance ranking (most important features in descending order). Panels (b) to (d) show the dependency plots between *tp_prob_1* and, respectively, *LAI*, *sdfor*, and *swvl*. Panels (e) to (g) show the same, but for *tp_prob_50*.

interactions of the rainfall-related variables with other features (expressed through colour gradients) demonstrate that environmental conditions modify the contributions of rainfall in predicting areas at risk of flash floods. Areas with dense vegetation coverage (higher *LAI* values, Figures 6.19b and e), primarily flatter (lower *sdfor* values, Figures 6.19c and f), and with mostly saturated soils (higher *swvl values*, Figures 6.19d and g) enhance sensitivity to rainfall, meaning that lower values of *tp_prob_1* are required to trigger higher (positive) SHAP contributions in the probability of having a flash flood in a grid-box.

6.5 Case Study: Storm Ida

For comprehensive details on Ida's synoptic history, rainfall patterns, and impacts, please refer to Section 4.5 in Chapter 4.

The model used for the case study is the same as that used in Sections 6.3.3 and 6.4. Figure 6.20 illustrates the estimates of areas at risk of flash flood with reanalysis data (Figure 6.20a) and the evolution of the predictions up to day 5 forecasts (Figures 6.20b-f). Two different events are highlighted: one due to a small-scale convective system (with the red circle in Figure 6.20a) and one due to a large convective system, i.e. Storm Ida (with the green circle).

Storm Ida: coherent flash flood risk signal with probabilities >10% maintained up to day 5, offering 120 hours lead time

For the Ida event, the Storm Event Database recorded over 50 flash flood reports across New Jersey, Pennsylvania, and New York on 1–2 September 2021, with particularly severe impacts in the New York City metropolitan area. The estimates of areas at risk of flash floods computed using the reanalysis data successfully identify the regions where these impacts were reported, with probabilities exceeding 10% across a spatially coherent area extending from Pennsylvania through New Jersey and into southern New York. Forecasts from day 1 to 5 (Figure 6.20b-f) show consistently elevated probabilities (>10%) up to medium-range lead times, providing a potential lead time of approximately 120 hours for preparedness actions.

Isolated convection yields lower, fragmented probabilities; predictability limited to day 1–3, consistent with NWP constraints

The isolated convective activity in Western CONUS presents a contrasting scenario. These smaller-scale systems, lacking the organised structure and predictability of a post-tropical cyclone, produce flash flood probabilities that are notably lower (1–3%) and more spatially fragmented. The model begins to indicate elevated probabilities over this region from

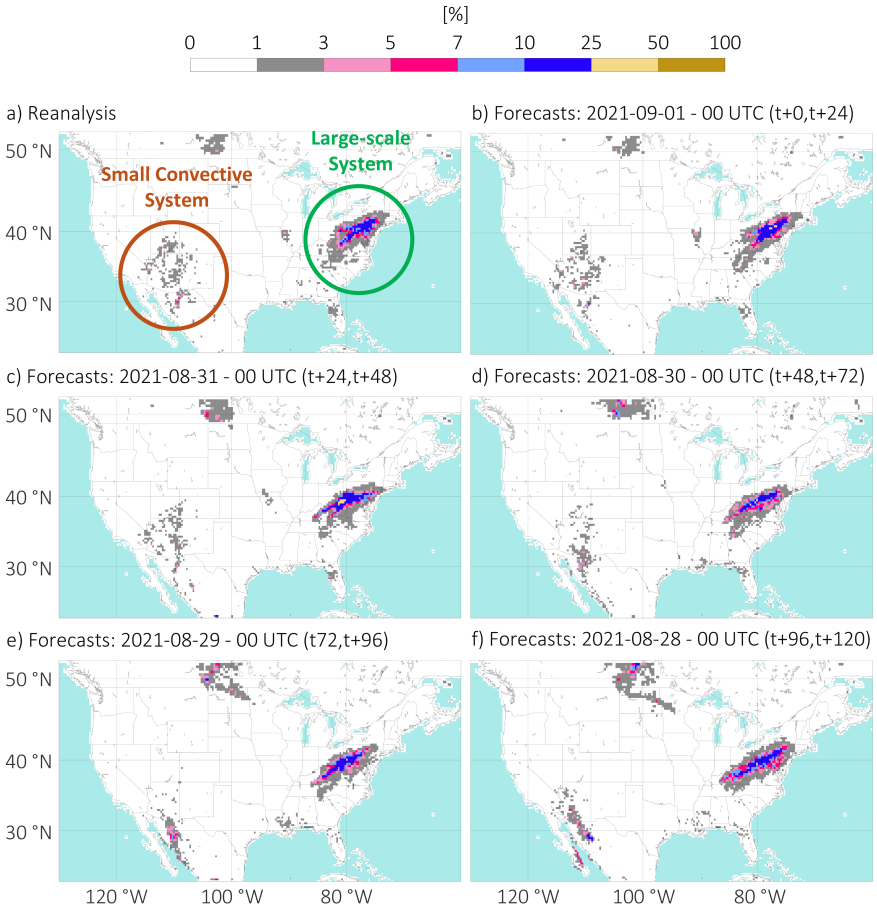

Figure 6.20: Areas at risk of flash floods. Probability of having a flash flood in a grid-box, valid for the 24-hourly accumulation ending on 2021-09-02 at 00 UTC. Panel (a) shows the probabilities computed with reanalysis data, while panels (b) to (f) show the probabilities computed with forecast data, respectively for day 1 (t+0,t+24), day 2 (t+24,48), day 3 (t+48,t+72), day 4 (t+72,t+96), and (t+96,t+120).

day 3 onwards (Figure 6.20d), but the signal remains diffuse compared to the Ida event. This difference illustrates a fundamental limitation: whilst the data-driven model can skilfully predict flash flood risk from large-scale, dynamically forced precipitation systems up to 5 days in advance, the prediction of flash floods triggered by isolated convection remains challenging beyond day 1 to 3 lead times. This limitation is consistent with the known predictability constraints of convective-scale phenomena in global NWP systems.

6.6 Discussions

The research presented in this chapter provides several insights that enhance the understanding of both the predictive capabilities and operational limitations of machine learning approaches in predicting areas at risk of flash floods.

Insights from model training

The comprehensive model training - carried out through the nested stratified cross-validation approach - and the comparison of different data-driven models provide interesting insights into the required model complexity to achieve good performance in predicting areas at risk of flash flood and how the use of severely imbalanced observational datasets may affect model training and forecast performance. Despite theoretical advantages in capturing non-linear relationships, the feed-forward neural network demonstrates no systematic superiority over tree-based methods, whilst requiring approximately 40 times longer training time. Gradient boosting implementations, particularly XGBoost, emerge as optimal choices, combining competitive performance with computational efficiency and interpretability. The hyperparameter importance analysis reveals that model performance depends primarily on fundamental architectural choices rather than sophisticated optimisation techniques. For gradient boosting methods, maximum tree depth and learning rate consistently dominate, whilst neural networks show the highest sensitivity to first-layer configuration. This finding suggests that, for example, if neural networks must be used, shallow neural networks might be preferred over deeper architectures, as the former is already able to capture the fundamental relationships between predictors and observations, thereby successfully identifying areas at risk of flash floods. In the case of gradient boosting, practitioners should prioritise careful selection of core architectural parameters over extensive hyperparameter search spaces. Finally, the comparative evaluation of balanced versus weighted loss functions reveals fundamental trade-offs in rare event prediction. Models employing weighted binary cross-entropy successfully enhance the identification of areas at risk of flash floods, achieving hit rates approaching 90% for gradient boosting implementations (compared to 40% in the case of using loss functions designed for more balanced datasets). However, this enhanced detection comes at the substantial cost of false alarm rates exceeding 50%, representing a twenty-fold increase compared to balanced configurations. This dramatic increase in false alarms has profound implications for operational deployment. Emergency management systems must strike a balance between the imperative to protect lives

through early warning and the risk of warning fatigue from excessive false alarms. These results suggest that for flash flood prediction, where public trust and response compliance are critical, the conservative approach of balanced loss functions may be preferable despite lower identification rates. The frequency bias and reliability diagrams support this conclusion, showing that balanced models provide more reliable predictions.

Model performance characteristics

The evaluation of six distinct machine learning architectures reveals good overall discrimination ability - with all models achieving AUC-ROC values clustering around 0.8 - and acceptable reliability - with all models achieving good reliability for probabilities less than 20% and overall frequency bias around 1. The AUC-PR metric reveals more nuanced performance variations, with values ranging from 0.02 to 0.03 across most models. Whilst these values appear modest, they represent meaningful skill when put into the context of the climatological frequency of flash floods in the observational dataset, i.e., 0.27% (0.0027). The precision-recall analysis exposes critical differences between architectures, with gradient boosting implementations and the feed-forward neural network maintaining higher precision at low recall thresholds compared to random forests, which exhibit delayed recall onset requiring extreme threshold relaxation to achieve any sensitivity. The minimal degradation between training (2001-2020) and verification (2021-2024) datasets also demonstrates a good model generalisation, indicating that the learned patterns capture fundamental hydro-meteorological relationships rather than data-specific anomalies.

Performance evolution from reanalysis to medium-range forecasts

The analysis of forecast skill degradation of estimates of probabilities of having a flash flood in a grid-box computed with reanalysis data to medium-range predictions (from day 1 to day 5) shows the predictability limits of the data-driven predictions of probabilities of having a flash flood event in a specific area. The discrimination ability, as estimated by the ROC curves and the AUC-ROC, shows a minimal degradation over time - from $\sim$0.8 to $\sim$0.75. On the contrary, the AUC-PR exhibits steeper degradation from $\sim$0.025 to $\sim$0.01, reflecting the increasing difficulty for the forecasts in maintaining precision for rare event detection as forecast uncertainty compounds. The reliability diagrams confirm that while low-probability predictions (<10%) maintain reasonable reliability at short lead times (for reanalysis data and day 1 forecasts), this threshold reduces considerably (<2%) beyond day 2.

Physical interpretation of data-driven predictions: (1) LAI may proxy seasonality due to correlation and not causality with flash flood occurrence, (2) soil moisture effects are physically consistent, and (3) orography results likely reflect impact reporting biases

The SHAP analysis establishes a clear hierarchy of predictive features that generally aligns with hydrological understanding. The overwhelming dominance of the rainfall parameter (*tp_prob_1*), accounting for 80% of mean absolute SHAP values, confirms that rainfall remains the primary driver of flash flood occurrence. The secondary features demonstrate modulating effects on the influence that the rainfall parameter has in enhancing the probability of a flash flood event. However, there is no complete alignment with the general hydrological understanding of their impact on flash flood occurrence. The vegetation coverage parameter (*LAI*) emerges as the second most influential parameter in determining the probability of a flash flood event. In flash flood susceptibility studies, vegetation parameters do not rank as high as in this study. Moreover, such parameters suggest that it is the lack of vegetation that increases the risk of flash flood occurrence. In contrast, these results indicate that the probability of having a flash flood increases with large values of LAI, which suggests dense vegetation. This odd correlation might be due to the fact that, climatologically, LAI values are greater over the summer (Owens and Hewson, 2018). The Storm Event Database indicates that, climatologically, flash flood events occur primarily during the summer, as this is when most convective flash-flood-triggering rainfall events occur (Davis, 2001). Thus this unusual behaviour regarding the vegetation coverage parameter may be due to the relationship between its highest values occurring over the same season as the highest frequency of flash floods occurs. More investigation is, therefore, needed to confirm this explanation, and if it is correct, engineer or substitute the LAI parameter to represent only vegetation coverage, with no secondary relationships to rainfall to avoid double-counting. Whilst incorporating time of year as a direct input would be inappropriate for a global operational system (given the hemispheric reversal of seasons), a controlled experiment over the CONUS, including temporal indicators, could help disentangle whether LAI is genuinely contributing predictive information about land surface conditions or merely acting as a proxy for the seasonal cycle of convective activity. After the vegetation coverage, the most important parameters influencing the probabilities of having a flash flood event are topographic steepness (*sdfor*) and antecedent soil moisture (*swvl*). This ranking agrees with flash flood susceptibility studies. The analysis of SHAP values suggests that high water content in the soil does modulate the effect of rainfall in defining the probabilities of having a flash flood event, i.e., the same amount of rainfall over a saturated soil increases the chances of having a flash flood event, compared to the case where the same rainfall amount falls over dry soil. This agrees with our understanding of the modulating effects of soil water content

on the effects that different rainfall amounts might have in the generation of flash flood events. On the contrary, the results provided by the SHAP values regarding the steepness of the orography provide the opposite of what physical hydrology suggests. The analysis of SHAP values suggests that primarily small values of sdfor (flatter areas) increase the probability of having a flash flood event. This contrasting result with physical hydrology may be due to where impact flash flood reports tend to be located. While it is known that flood water rises rapidly in complex orographic areas, the majority of the impacts are seen in the valleys, downstream of the complex orographic areas, where typically most of the people live and where the majority of the impacts are reported. It is, therefore, assumed that this odd relationship with the orographic parameters is the result of the type of observational dataset considered in this analysis (impact reports) rather than a genuine physical relationship, and that this result might change if discharge observations were considered instead.

Considerations for operational deployment

Based on the comprehensive evaluation across training, verification, and case study analyses, the XGBoost implementation of gradient boosting trained with balanced loss function and optimised for AUC-ROC emerges as the recommended configuration for operational deployment. This recommendation rests on multiple converging lines of evidence that address both technical performance and practical constraints. The selected configuration maintains stable AUC-ROC around 0.8 and high AUC-PR values (from $\sim$0.03 and 0.01) across lead times. The XGBoost implementation of gradient boosting also shows frequency bias closest to 1 when probabilities are computed with reanalysis data and stable around 2.5 when forecasts up to medium-range lead times are considered. It is also the architecture that maintains good reliability for the highest probability threshold when considering reanalysis data (< 10%) and forecasts (< 2%). The case study supports this choice, demonstrating a good correspondence in XGBoost between predictions of areas at risk of flash floods and observations, without the noisy patterns exhibited in other models (not shown), which produce a high number of grid-boxes with probabilities between 1 and 3% in areas where there were no records of flash flood events. Finally, from a computational perspective, XGBoost's training time of approximately 20 minutes per fold enables regular model updates as new event reports become available. LightGBM and CatBoost also appear as good options. However, CatBoost has a significantly longer training time than XGBoost and LightGBM (i.e., 2 hours), which may make it less suitable for operational applications.

6.7 Conclusions

This chapter has demonstrated the feasibility of developing data-driven predictions of flash flood risk at regional scales using hydro-meteorological reanalysis data and impact reports, successfully addressing Research Question 2 of this thesis.

XGBoost achieves optimal performance; shallow neural networks comparable but 40 times more costly to train

The comprehensive evaluation of six machine learning architectures reveals that gradient boosting implementations, particularly XGBoost, achieve optimal performance in predicting areas at risk of flash floods over the CONUS. Shallow neural networks achieve a similar performance but at a 40 times higher training cost (8 hours rather than 20 minutes), which may make it less palatable for operational implementations.

Hydrological and static features critically modulate the dominant rainfall signal in flash flood prediction

The integration of hydrological features (e.g., soil moisture, vegetation indices, and topographic characteristics) beyond only precipitation variables (as seen in Chapter 5) enhances predictive capabilities. While the SHAP analysis confirms that 1-hour precipitation probability dominates model decisions with an 80% contribution, hydrological, climatological, and static features critically modulate the impact of the rainfall variables in predicting areas at risk of flash flood.

Skill degradation informs tiered operational protocols: protective actions at short range, preparedness at medium range

Skill degradation analysis establishes clear operational protocols. High-confidence warnings between 0-24 hour forecasts may be used for taking (high-cost) protective actions such as evacuation of people, while less confident forecasts (lead times > t+24) may be used for (low-cost) preparedness actions such as moving belongings to higher floors or strategic planning such as calling people to monitor possible flash flooding situation and be ready to take actions.

Future work: higher-resolution NWP integration, ensemble probability estimation, and assessment of global transferability

While this development uses rainfall forecasts post-processed with the ecPoint methodology (which produces rainfall estimates at point-scale), future research should prioritise integration with higher-resolution global NWP models (e.g. IFS at 9 km) or with convection-permitting models (e.g. Destination Earth implementations at 4 km resolution) to better capture localised hydro-meteorological processes, addressing current limitations of 31-kilometre grids. While this study provides a single probability estimate per grid box for each model, future work could generate an ensemble of probability values by combining predictions from multiple data-driven architectures through techniques such as bagging or stacking. Such an ensemble approach would provide uncertainty bounds on the flash flood

probability estimates, enabling forecasters to distinguish between situations where models agree (high confidence) and those where they diverge (low confidence)—information that is essential for operational decision-making. The models developed here are valid over the CONUS, but could potentially be run operationally using global NWP model output and produce predictions over a continuous global domain. However, the best approach to do this remains to be assessed.

CHAPTER 7

TOWARDS PREDICTIONS OVER A CONTINUOUS GLOBAL DOMAIN: GLOBAL IMPLEMENTATION OF REGIONALLY-TRAINED MODELS

7.1 Introduction

Inequitable observational infrastructure creates urgent need for global flash flood prediction approaches

Developing continuous global flash flood predictions is not merely a technical challenge; it also raises critical questions about equity in disaster risk reduction, given that prediction skill and observational infrastructure are unevenly distributed across regions of differing vulnerability. Flash floods represent one of the most devastating natural hazards globally, affecting populations in the Global North and Global South equally (Dordevic et al., 2020; Yin et al., 2023). However, the observational infrastructure necessary for developing predictive models that can help the population prepare and mitigate the risk against flash floods remains concentrated in a small subset of wealthy nations (Kratzert et al., 2023). This disparity violates the principle that all populations, regardless of their location or economic status, should have access to life-saving flood warnings — a goal central to the UN's 'Early Warnings for All' initiative. Moreover, as flash floods typically occur in poorly gauged catchments, this further reduces the number of observations available for model development and post-hoc event analysis even in data-rich regions in the Global North, thereby hindering our understanding of flash flood generation mechanisms (Gaume et al., 2009, 2016). This inequitable distribution of observational capacity to assess flash flood occurrence creates an urgent need for innovative approaches

that can transcend the limitations of traditional catchment-specific modelling paradigms to develop predictive models over a continuous domain able to truly cover all populations around the globe.

Data-driven flash flood prediction remains largely catchment-scale due to observation scarcity

The emergence of data-driven approaches in large-sample hydrology has demonstrated remarkable success in learning complex relationships between hydro-meteorological variables and flood occurrence (Nearing et al., 2024). Many data-driven applications are also now applied to predict areas at risk of flash floods or river discharge in flashy catchments (Oddo et al., 2024; Zhao et al., 2025). However, such applications remain primarily at the catchment level due to the aforementioned severe paucity of observations suitable for predicting flash flood events. There exist only a few examples of prediction systems at a larger scale (Liu et al., 2018)[1].

Transfer learning from data-rich regions offers a pathway to continuous global flash flood prediction

Recent advances in transfer learning and domain adaptation offer a transformative approach to this challenge. Rather than requiring comprehensive local observations for model training, one could train a data-driven model to learn generalisable hydro-meteorological relationships from data-rich regions and subsequently deploy the model to create predictions over a continuous global domain, provided that hydro-meteorological variables from global NWP models are used (Gupta et al., 2014; Kratzert et al., 2024). The key insight underlying this approach is that whilst specific catchment characteristics may vary globally, the fundamental physical processes governing flash flood generation — the interaction between intense precipitation, antecedent soil moisture (affecting infiltration rates), topography, and land surface characteristics — exhibit sufficient commonality to enable knowledge transfer across different regions.

Fundamental trade-off: spatial coverage versus observational density in transferable model development

The development of such transferable models faces a critical trade-off between spatial coverage and data density. Models trained on high-density observations from limited geographical regions may capture local flash flood dynamics with high fidelity, but may fail to generalise to regions with different climatic regimes, topographies, or land use patterns. Conversely, models trained on sparse global datasets may achieve broader applicability but at the cost of reduced accuracy in the overall identification of areas at risk of flash floods.

---

[1]In the coming years, researchers at Karlsruhe Institute of Technology will also develop a new data-driven flash flood forecasting system over the whole of Germany

7.2 Data

The data used in this chapter has already been presented in Chapter 4 and Chapter 6. The target variable - i.e., flash flood impact reports from the Storm Event Database - are described in Section 4.2. The hydrological (antecedent soil moisture), climatological (vegetation coverage), and static (orography slope) feature variables are described in Section 4.3. Finally, the rainfall variables used for the development of the data-driven models are described in Section 4.4.

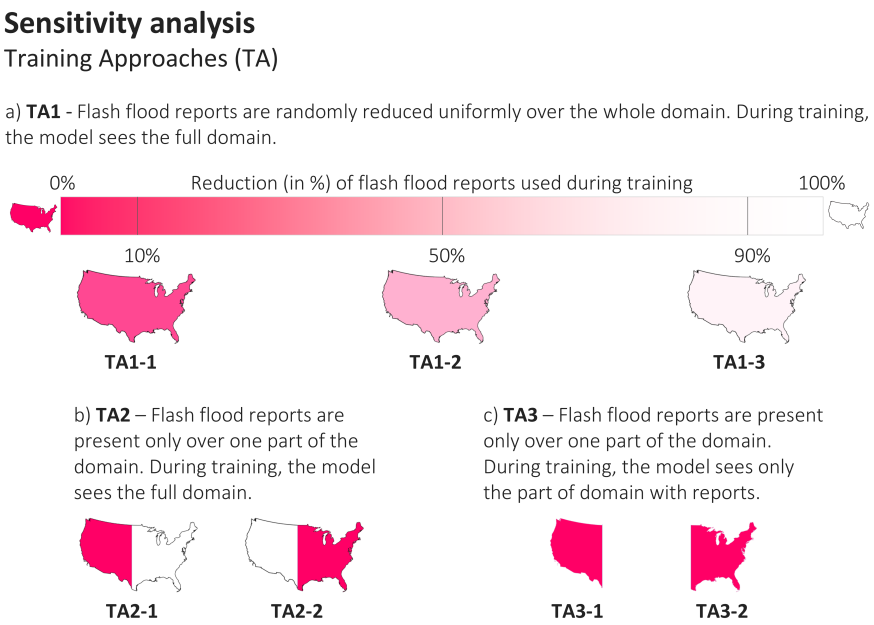

Figure 7.1: Training approaches adopted in the sensitivity analysis. Panel (a) describes Training Approach 1 (TA1) - where the flash flood reports are randomly reduced uniformly over the whole domain by 10% (TA1-1), 50% (TA1-2), and 90% (TA1-3), and during training, the model sees the full domain. Panel (b) describes Training Approach 2 (TA2) - where the flash flood reports are present only over one part of the domain (TA2-1 for reports in the west and TA2-2 in the East), but during training, the model still sees the full domain. Panel (c) describes Training Approach 3 (TA3) - where the flash flood reports are present only over one part of the domain (TA3-1 for reports in the west and TA3-2 in the East), and during training, the model sees only the part of the domain with reports.

7.3 Methods

7.3.1 Training approaches

The methodological framework examines three distinct training strategies to determine the optimal approach for developing flash flood predictions

across a continuous global domain, considering heterogeneous data availability.

Training approach 1 (TA1)

This approach tests whether training over the full CONUS domain remains effective when flash flood observations are systematically reduced. Flash flood reports from the Storm Event Database were randomly sampled at three levels: 90%, 50%, and 10% of the original dataset. The sampling was applied uniformly across the entire domain, maintaining the spatial distribution whilst reducing observation density. This approach simulates the scenario of training a global model with sparse but spatially distributed observations.

Training approach 2 (TA2)

The second approach maintains training over the whole CONUS domain but restricts flash flood observations to specific regions. The model receives hydro-meteorological data from the entire domain during training, but only has access to flash flood labels in the selected region. The non-selected regions contribute only negative samples (non-flood events) to the training dataset. This configuration simulates the realistic scenario of developing a global model where flash flood reports are available only from certain countries or regions, whilst meteorological data has global coverage.

Training approach 3 (TA3)

The third approach examines whether models trained on high-quality observations from limited geographical regions can effectively predict flash floods in areas where they have never observed any events. The CONUS domain was divided into four regions: East (east of 98°W), West (west of 98°W), North (north of 37°N), and South (south of 37°N). For each configuration, the model was trained using flash flood observations from only one region, with the remaining regions containing no observations during training. The trained model was then applied to predict flash floods across the entire CONUS domain. This approach directly addresses the scenario where certain parts of the world have excellent observational infrastructure, whilst others have none.

East-west spatial division at 100°W separates contrasting flash flood regimes by topography, population, and storm type

The spatial divisions for Approaches 2 (Figure 7.1b) and 3 Figure 7.1c were selected to create regions with varying flash flood characteristics. Two main regions were selected. The eastern region (east of the longitude 100°W) encompasses mostly flat regions (except for the Appalachian Mountains), is highly populated, and is frequently affected by convective storms and hurricanes. The western region (west of the longitude 100°W) encompasses mostly mountainous regions (except for the Central-North Pacific

coast), is less populated, and is more frequently affected by large-scale systems bringing large amounts of rainfall due to orographic enhancement.

7.3.2 Training data configuration

Baseline configuration: same features, observational dataset, and training/verification periods as previous chapters

For all training approaches, the baseline configuration consists of the full Storm Event Database over the CONUS as presented in Chapter 4. The hydro-meteorological features - including ERA5-ecPoint rainfall, ERA5 antecedent soil moisture, vegetation coverage and topography steepness - also remain consistent with those presented in Chapter 4 and already used in Chapter 6. The training period remains between 2001 and 2020, as well as the verification period from 2021 to 2024.

7.3.3 Model configuration

XGBoost hyperparameters fixed to isolate effects of training data modifications on performance

The analysis employs the XGBoost model developed in Chapter 6. The model's hyperparameters remain consistent with those optimised in Chapter 6 to ensure that performance differences arise solely from training data modifications (exemplified by TA1-3) rather than model architecture changes.

7.3.4 Performance evaluation

Objective verification using same scores as previous chapter to quantify performance degradation across training approaches

Following the approach adopted in Chapter 6, an objective verification analysis will be carried out for the three TAs over the full CONUS domain, and the results will be compared to those obtained in Chapter 6 using the complete training dataset. Such a comparison will determine any potential degradation in predictive performance for any of the three TAs. Hence, the same verification scores adopted in Chapter 6 will also be considered in this chapter, i.e. ROC curves, precision-recall curves, and reliability diagrams as breakdown scores. Area under the ROC (AUC-ROC) and under the precision-recall curves (AUC-PR), together with the frequency bias (FB) will also be considered as overall scores for discrimination ability (AUC-ROC and AUC-PR) and reliability (FB).

Case studies supplement objective verification where global impact databases lack density: Storm Ida, Valencia 2024, and China 2021

The verification assessments will conclude with a case-study-based analysis. This type of *subjective* analysis is preferred to analyse the performance of the global predictions because global impact datasets, such as EM-DAT and DesInventar, do not contain enough observations to compute robust statistics from objective verification (as done over the CONUS). For

the US, predictions of areas at risk of flash floods obtained from the three TAs will be presented again for Storm Ida (for more details on the specific event, refer to Chapter 4). Moreover, this US-focused case study will be complemented in this chapter by case studies from other parts of the world: the flash floods in Valencia, Spain, in 2024 and in China, in 2021. Both represent extremely severe flash flood events that caused high death tolls and considerable economic losses. However, the first will represent a very localised flash flood event, generated by a small-scale convective system. In contrast, the second will represent a widespread flash flood event, generated by a large-scale convective system. These two events were chosen due to their inherent differences in predictability.

7.4 Results

Verification results – Overall scores

Evaluated over the **verification** dataset, for the XGBoost implementation of gradient boosting

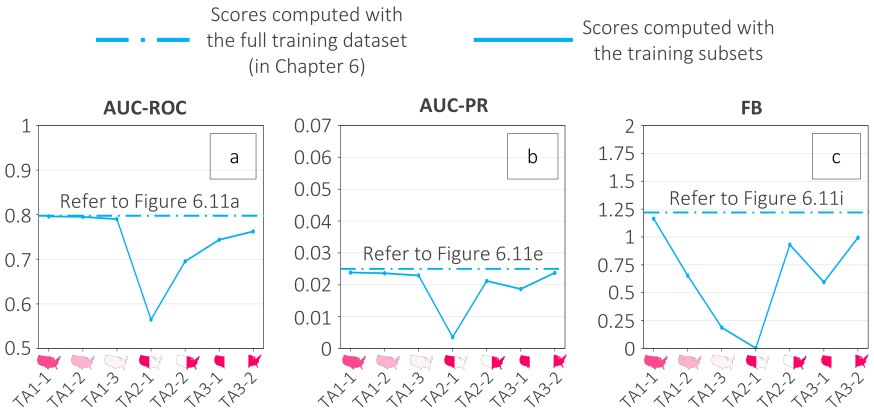

Figure 7.2: Objective verification: overall scores (AUC-ROC, AUC-PR, FB) Panel (a) shows the area under the ROC curve (AUC-ROC), computed for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets and hyperparameters optimised to maximise AUC-ROC (as developed in Chapter 6). The score is computed with data from the verification dataset (from 2021 to 2024). The solid line shows the scores computed using the three considered training approaches (TA1-3) as described in Figure 7.1. The dashed line represents the score obtained when the model was trained with the full training dataset (refer to Figure 6.14a. Panels (b) and (c) show, respectively, the area under the precision-recall curve and the frequency bias. Refer to Figures 6.14e and 6.14i for the corresponding scores obtained when the model was trained with the full training dataset.

Objective verification: overall scores (AUC-ROC, AUC-PR, FB)

The area under the ROC curve (AUC-ROC, Figure 7.2a) and the area under the precision-recall curve (AUC-PR, Figure 7.2b) show that TA1 and TA3 achieve a very close discrimination ability to that achieved by the model trained with the complete training dataset (dash lines). TA2 is the approach

that reduces the most the discrimination ability. In particular, TA2-1 reduces both AUC-ROC and AUC-PR by $\sim$20% (from 0.8 to 0.57 and from 0.025 to 0.0024). For TA1, the frequency bias (FB) diminishes with increasing reductions of flash flood reports in the training dataset. The expectation of FB scaling proportionally with the reduction in training reports is easily verifiable by the FB values for TA1-1 to TA1-3 in Figure 7.2c, which reduce proportionally to the reduction of impact reports considered during training: given a FB=1.2 for TA1-1, the FB values for TA1-2 and TA1-3 would be expected to be, respectively, $\sim$0.66 and $\sim$0.13. These values are very close to those computed, 0.6 and 0.15, respectively. The datasets in TA2 and TA3 show opposite behaviours. TA2-1 and TA3-1 (which consider reports only over the West part of the CONUS) show poorer FBs with an overall tendency to underestimate the occurrence of flash floods. In particular, TA2-1 shows an extremely low FB = 0.0022. TA2-2 and TA3-2 (which consider reports only over the East part of the CONUS) show the closest values to those obtained for the model trained with the complete training dataset, even though with a slight tendency to underestimate the frequency of flash floods.

Objective verification: breakdown scores (ROC curves)

When considering the ROC curves for TA1, those computed with probability thresholds discretised every 0.01% (dashed lines in Figures 7.3b-c) show negligible difference in shape to that computed with the full training dataset (dashed line in Figure 7.3a). However, when considering the ROC curves built with probability thresholds discretised every 1% (solid lines in Figures 7.3b-d), the ROC curves get ”squashed” towards the bottom-left corner of the unit square. Consequently, the values of AUC-ROC reduce from 0.652 for TA1-1 to values close to the ”no-skill” threshold (equal to 0.5) for TA1-2 (0.588) and TA1-3 (0.512). As seen in Figure 7.2a, the training datasets in TA2 and TA3 show opposite discrimination abilities. TA2-1 (Figure 7.3e) and TA3-1 (Figure 7.3g), with flash flood reports in the West part of the CONUS, show the poorest ROC curves. In particular, the ROC curve for TA2-1 shows that the model trained with this dataset has no discrimination ability. In contrast, the models trained with the TA2-2 (Figure 7.3f) and TA3-2 (Figure 7.3h) datasets show the closest ROC curves (with both discretisations) to the one trained with the complete training dataset (Figure 7.3a). The only difference lies in the shape of the dashed ROC curves for TA2-2 and TA3-2. For very small probabilities of exceedance (< 0.1%), the ROC curves are squashed to the diagonal of the unit square, showing that the model has no discriminative ability when forecasts produce very small probabilities of exceedance.

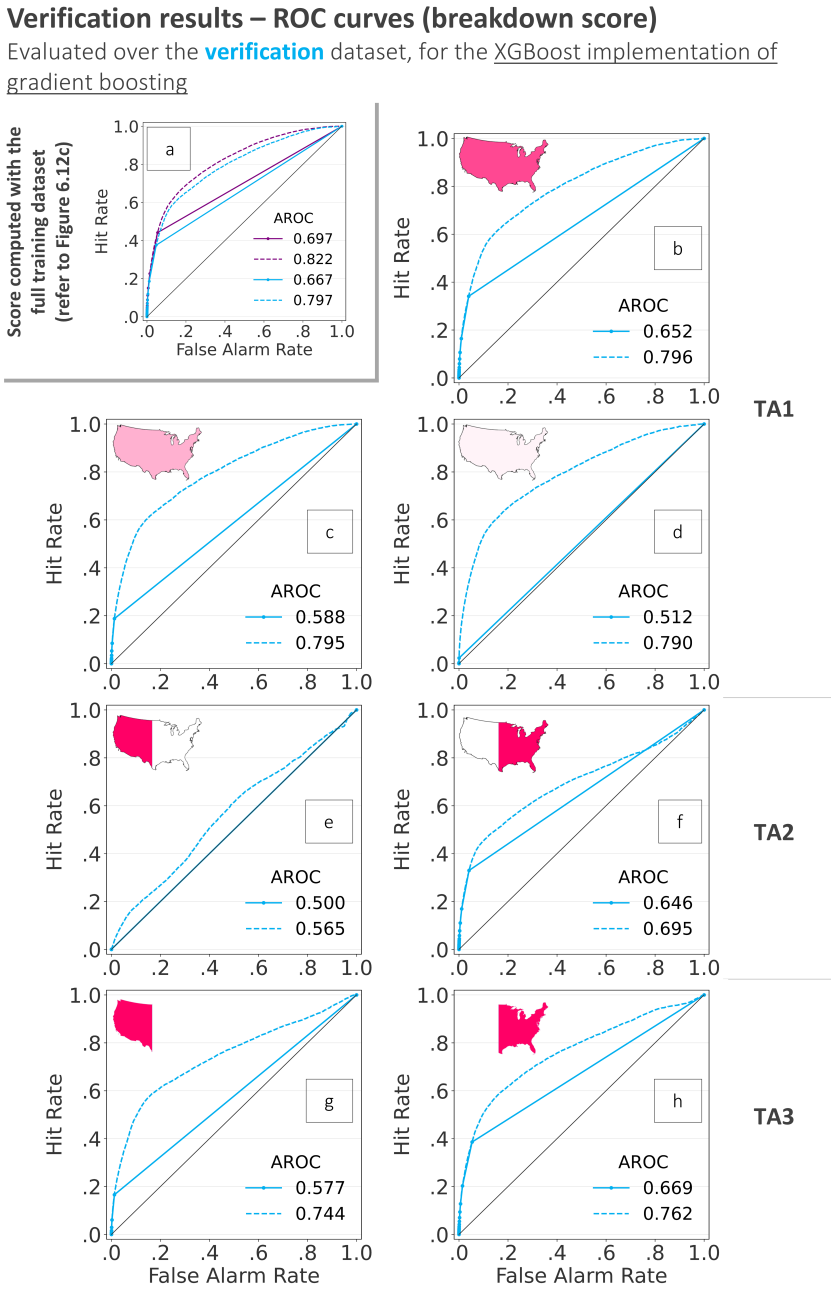

Figure 7.3: Objective verification: breakdown scores (ROC curves) ROC curves are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.15c). Panel (a) shows the ROC curve for the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the ROC curves for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the ROC curves obtained for TA2, while panels (g) and (h) show the ROC curves obtained for TA3. The ROC curves drawn with solid lines correspond to a probability discretisation of 1%, with probabilities of exceedance ranging from 0 to 99%. The ROC curves drawn with dashed lines correspond to a probability discretisation of 0.01%, with probabilities of exceedance ranging up to values that depend on the TA, and shown in Figure 7.1.

Objective verification: breakdown scores (precision-recall curves)

All the precision-recall curves have a similar shape to that obtained by the model trained with the full training dataset (Figure 7.3a). For TA1 (Figure 7.3b-c), the precision at very small values of recall increases with increasing reductions of flash flood reports in the training datasets. In contrast, the values of recall produced by the models decrease. Again, TA2-1 (Figure 7.3e) and TA3-1 (Figure 7.3g) correspond to the training approaches showing the worst performance. Both are capable of producing only very small values of recall, and TA2-1 also produces values of precision that are extremely low and close to 0. In contrast, TA2-2 (Figure 7.3f) and TA3-2 (Figure 7.3h) show the best precision-recall curves, achieving both high values of precision and recall.

Objective verification: breakdown scores (reliability diagrams)

The reliability diagrams for TA1-1 Figure 7.5b, TA2-2 Figure 7.5f, and TA3-2 Figure 7.5h show very similar reliability to that obtained by the model trained with the complete training dataset (Figure 7.5a): forecast probabilities below $\sim$20% remain fairly reliable (as they overlap with the diagonal), and the probabilities produced by the model remain between 30 and 40%. The forecast probabilities for TA1-2 (Figure 7.5c) remain overall reliable, but the model is not able to produce probabilities bigger than $\sim$20%. When considering TA1-3 (Figure 7.5d), the model is producing very small forecast probabilities that, overall, underestimate the observed frequency of flash floods (the diagram lies above the diagonal). The model trained with TA3-1 (Figure 7.5g) shows a similar behaviour to TA1-3. Finally, TA2-1 (Figure 7.5e) shows an almost inexistent reliability diagram, showing that the model is producing very small probabilities of flash flood occurrence.

7.5 Case Study over the CONUS: Storm Ida

For comprehensive details on Ida's synoptic history, rainfall patterns, and impacts, please refer to Section 4.5 in Chapter 4.

Storm Ida maps confirm verification results: TA1-1 and TA3-1 match baseline; spatial bias degrades TA2-2 westward; TA2-1 performs worst

The map plots for Storm Ida further illustrate the performance differences identified in the verification metrics for different TAs. The performance of TA1-1 (Figure 7.6b) and TA3-1 (Figure 7.6h) remain very similar (i.e., similar distribution of probabilities) across Western and Eastern CONUS to the baseline prediction of probabilities of areas at risk of flash floods, computed with the full training dataset (Figure 7.6a). Over the Eastern side of the CONUS, TA2-2 (Figure 7.6f) also provides very similar probabilities over the areas surrounding New York. However, the spatial bias of no flash flood reports seen by the model over the whole training

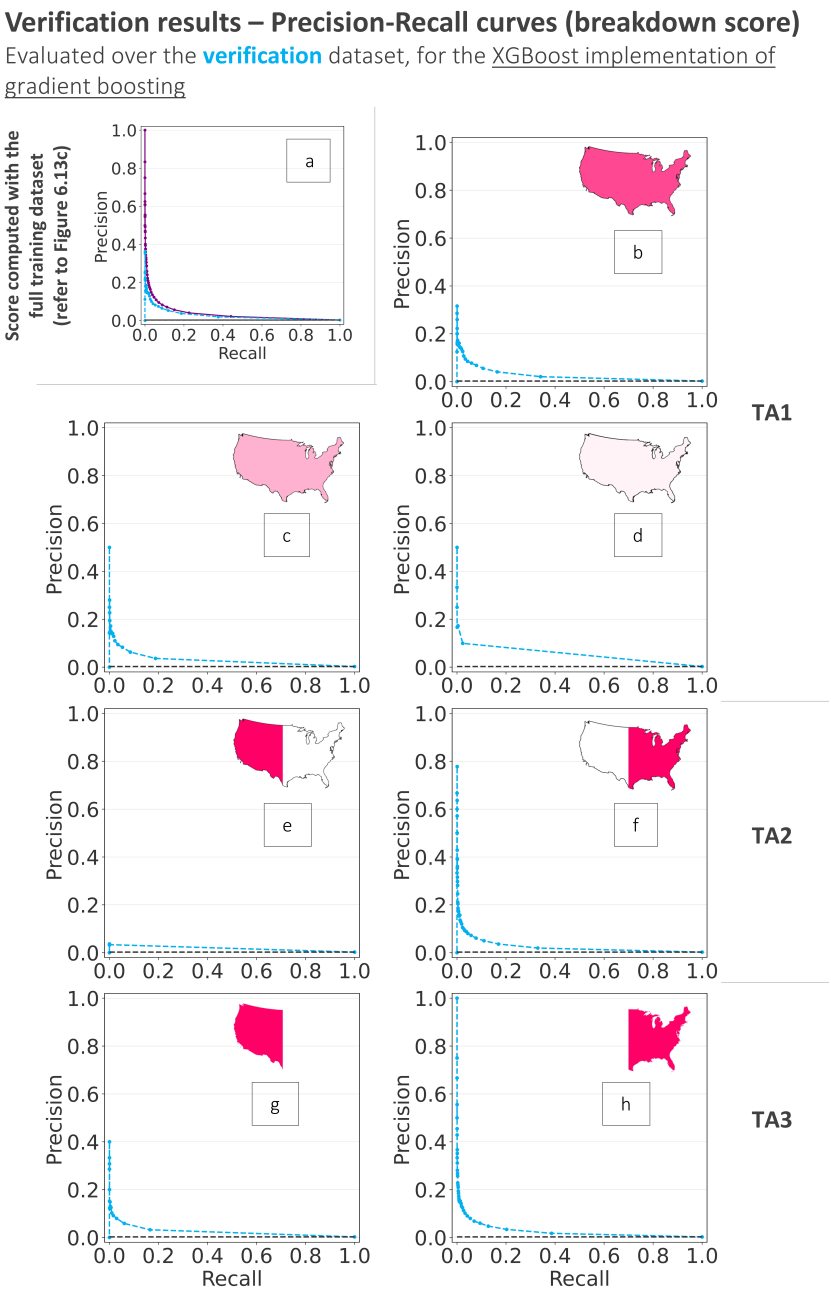

Figure 7.4: Objective verification: breakdown scores (Precision-Recall curves) Precision-Recall curves are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.16c). Panel (a) shows the precision-recall curve for the model trained with the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the precision-recall curves for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the ROC curves obtained for TA2, while panels (g) and (h) show the ROC curves obtained for TA3.

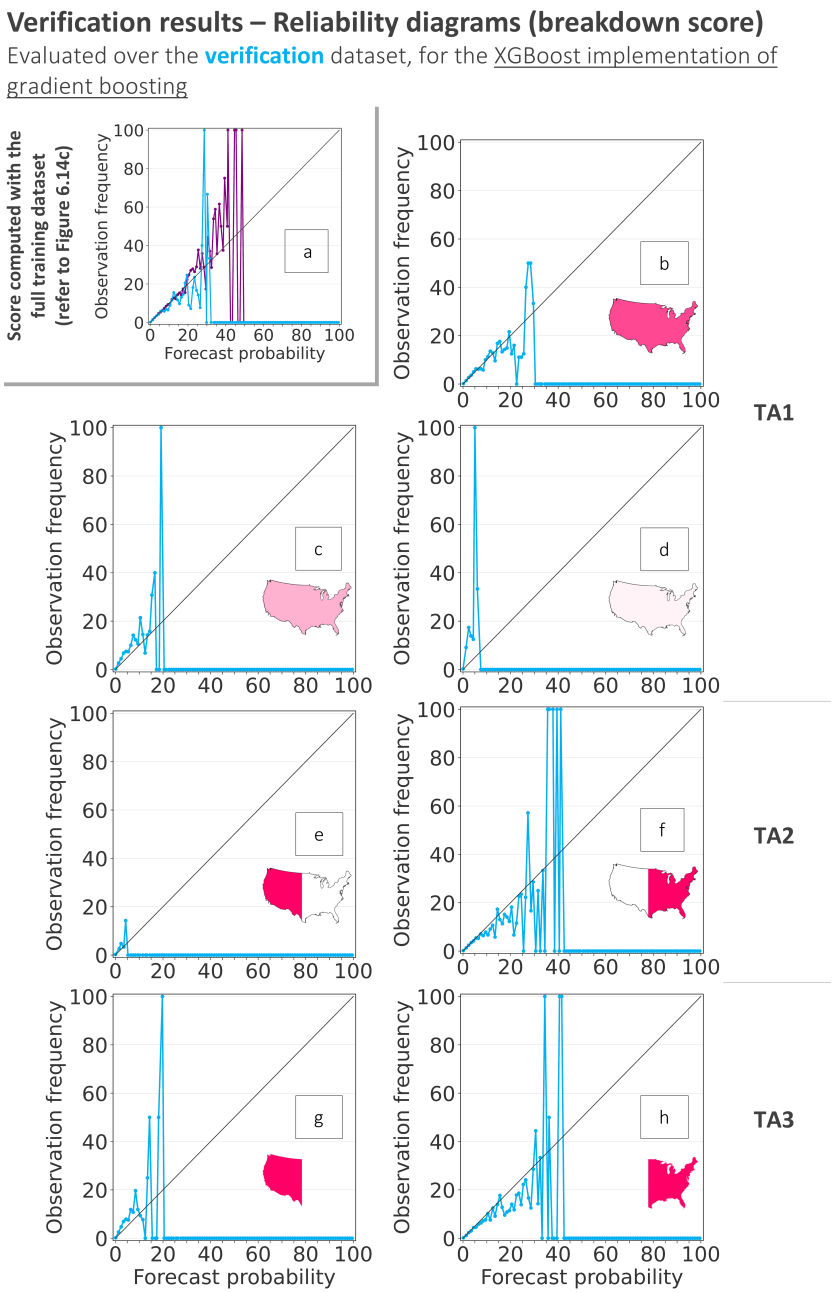

Figure 7.5: Objective verification: breakdown scores (Reliability Diagrams) Reliability diagrams are shown for the XGBoost implementation of gradient boosting, trained with the loss function for balanced datasets, and hyperparameters optimised to maximise AUC-ROC (developed in Chapter 6). All panels refer to reanalysis data (as in Figure 6.17c). Panel (a) shows the reliability diagram for the model trained with the training dataset - from 2001 to 2020 - and the verification dataset - from 2021 to 2024). Panels (b) to (d) show the reliability diagrams for the model trained with the training subset corresponding to training approach 1 (TA1), only for the verification dataset. Panels (e) and (f) show the reliability diagrams obtained for TA2, while panels (g) and (h) show the reliability diagrams obtained for TA3.

period causes it to reduce the probabilities of having a flash flood over the Western side close to zero. The training approaches TA1-2 (Figure 7.6c), TA1-3 (Figure 7.6d), TA2-1 (Figure 7.6e), and TA3-1 (Figure 7.6g) all demonstrate reduced flash flood probabilities across both Eastern and Western CONUS. Among these, TA2-1 (Figure 7.6e) exhibits the poorest performance, with probability values diminishing to near-zero values throughout the domain.

7.6 Case studies for regions outside the CONUS

Results from the previous section showed that TA3-2 achieved the overall best performance of all training approaches considered in the sensitivity analysis. Hence, for the case studies outside the CONUS, the probabilities for areas at risk of flash floods were computed with the XGBoost model configuration as described in Section **??**, with the model exposed only to the CONUS domain and to the impact reports from the Storm Event Database (which represents the application of TA3-2 to compute global probabilities).

7.6.1 Flash floods in Spain in October 2024

Flash floods in Spain in October 2024: event description

Spain's Mediterranean and adjacent regions (Albacete, Cuenca and Málaga) suffered a prolonged period of intense rainfall between the 28th of October and the 4th of November 2024. On the 29th of October, a cut-off low-pressure system (sometimes referred to as a "Dana" in Spanish) over the Strait of Gibraltar, favoured the formation of stationary low-pressure systems that brought moist air from the Mediterranean Sea towards the eastern coast of Spain over many hours that day[2]. Rainfall totals reached record-breaking totals for 1-hourly (184.6 mm), 6-hourly (620.6 mm), and 12-hourly (720.4 mm). In Turis Mas de Calabarra, 771.8 mm were observed in 24-hours (Figure 7.7a), a value second only to the 817.0 mm observed in Oliva (Valencia) in 1987. The torrential rainfall led to widespread severe surface

[2]For more detailed information about the event, please refer to the following AEMET report (in Spanish): https://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/estudios/informe_episodio_dana_29_oct_2024_.pdf

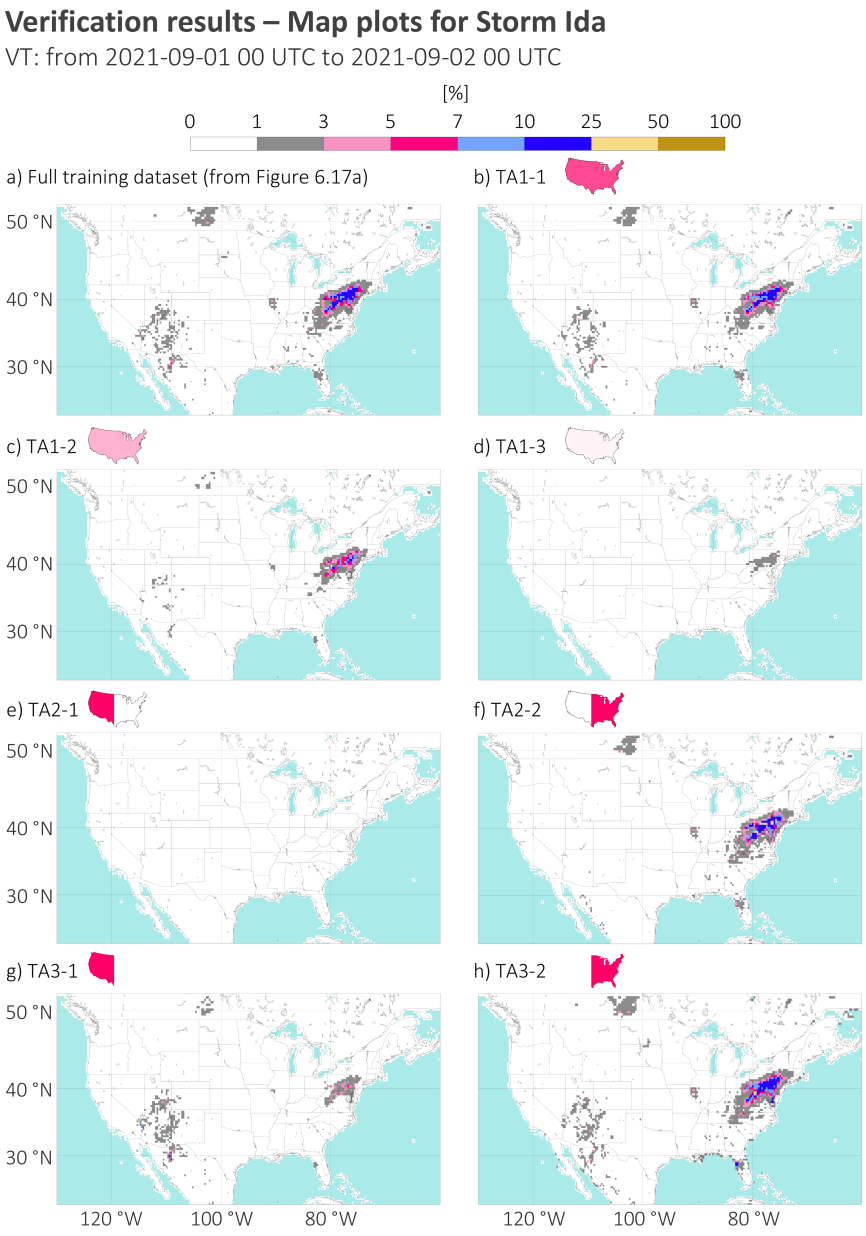

Figure 7.6: Map plots of probabilities for areas at risk of flash flood for different training approaches, for reanalysis data. Panel (a) shows the baseline probabilities obtained by training the considered XGBoost model (with balanced loss function and hyperparameters optimised by maximising the AUC-ROC metric) with the full training dataset as shown also in Figure 6.20a, in Chapter 6. Panel (b) to (d) show the probabilities of areas at risk of flash floods for TA1, respectively, TA1-1, TA1-2, and TA1-3. Panels (e) and (f) show the probabilities for TA2, respectively, TA2-1, and TA2-2. Panels (g) and (h) show the probabilities for TA3, respectively for TA3-1 and TA3-2.

runoff and riverine flash flooding over the province of Valencia. Most of the heaviest rainfall totals were observed over the small catchment (380 km^2) of the river Rambla del Poyo. Hence, the downstream city of Valencia was one of the most impacted areas despite receiving little direct precipitation. Overall, this event caused €16.5 billion in infrastructure damage and varying economic losses, thousands of displaced individuals, and 232 deaths.

Flash floods in Spain in October 2024: rainfall-based predictions of areas at risk of flash flood

The rainfall predictions for the day of the Valencia flash floods indicated that Spain would experience the heaviest precipitation across the entire Mediterranean region (Figures 7.7b-c). Based on rainfall climatology computed with 24-hourly ERA5-ecPoint rainfall estimates between 1991 and 2020, the 1-year return period corresponds to 40-50 mm/24h over the Valencia Region. Probabilities of exceeding this threshold reached 70% west of Valencia, particularly in the upstream area of the Rambla del Poyo catchment. The city of Valencia itself showed lower but still significant probabilities of up to 50%. The 50-year return period analysis, corresponding to 150-200 mm/24h, provided greater spatial precision in identifying areas that ultimately experienced severe rainfall, i.e., the Valencia Region, Cuenca, Albacete, and Malaga (Figure 7.7c). The signal up to day-5 forecasts was consistent about the areas that could have been affected by extreme rainfall, despite showing smaller probabilities at longer lead times (not shown).

Flash floods in Spain in October 2024: data-driven, hydro-meteorological predictions of areas at risk of flash flood

The data-driven model's spatial predictions of flash flood risk largely corresponded to the areas already identified by rainfall forecasts. The predicted probabilities of flash flood occurrence were modest in absolute terms (generally below 10%). However, the model's primary contribution does not lie in the absolute magnitude of probabilities but in their relative spatial distribution — specifically, its ability to propagate the flash flood risk signal downstream, to the city of Valencia, from the upstream areas of the Rambla del Poyo catchment. Rainfall forecasts indicated approximately 20% lower risk over Valencia compared to upstream areas. In contrast, the data-driven model maintained elevated probabilities over Valencia, recognising that extreme rainfall over the Rambla del Poyo catchment would generate comparable flash flood risk in the downstream urban area, despite Valencia receiving less direct rainfall. This enhanced capability stems from the model's consideration of rainfall probabilities in adjacent grid-boxes alongside those in the target location, effectively capturing the hydrological connectivity between upstream precipitation and downstream flooding. Consequently, the data-driven model reduced the difference in risk between inland areas and Valencia from 20% to approximately 5%, providing a more

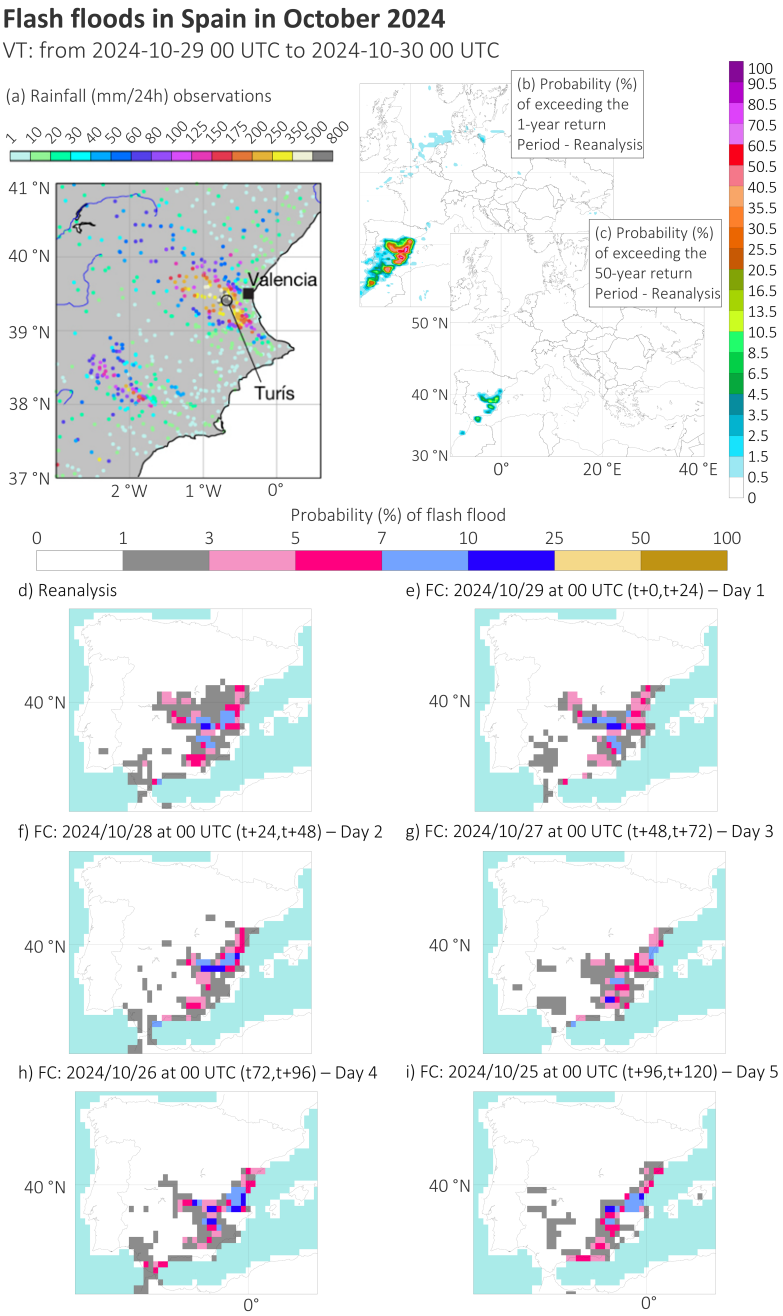

Figure 7.7: Flash floods in Spain in October 2024. The valid time (VT) for all plots is from the 29th of October 2024 at 00 UTC to the 30th of October 2024 at 00 UTC. Panel (a) shows rainfall observations (mm/24h) over Valencia, taken from Gascón et al. (2025). Panels (b) and (c) show the ERA5-ecPoint rainfall probabilities (%) of exceeding, respectively, the 1- and 50-year return period from reanalysis over Europe. Panel (d) show the probability (%) of flash flood computed with reanalysis data for a zoomed-in area over Spain. Panels (e) to (i) show the probability of flash flood computed from ERA5-ecPoint rainfall forecasts (FC), respectively, for day 1 - 2021/07/20 at 00 UTC (t+0, t+24) - day 2 - 2021/07/19 at 00 UTC (t+24, t+48) - day 3 - 2021/07/18 at 00 UTC (t+48, t+72) - day 4 - 2021/07/17 at 00 UTC (t+72, t+96) - and day 5 - 2021/07/16 at 00 UTC (t+96, t+120).

accurate representation of the flood hazard that ultimately materialised in the city.

7.6.2 Flash floods in China in July 2021

Flash floods in China in July 2021: event description

Between the 17th and the 20th of July 2021, the Henan Province in North-East China experienced extremely severe rainfall and widespread severe flash flooding. On the 20th of July, 624.1 mm/24h were recorded in the province's capital, Zhengzhou (Figure 7.8a, red circle), nearing the year's average precipitation. On the same day, the city also experienced the most intense rainfall - 201.9 mm/1h between 4 and 5 pm local time - ever recorded since measurements began in 1951. Overall, 8,150,00 people were evacuated, 14.5 million people were somehow affected around the province, and 398 people died.

Flash floods in China in July 2021: rainfall-based predictions of areas at risk of flash flood

The rainfall probabilities of exceeding the 1-year return period were computed from ERA5-ecPoint reanalysis (Figure 7.8b). These probabilities effectively highlighted the three wettest areas on the 20th of July, as observed over the region displayed in Figure 7.8a. The first area encompasses Henan's capital, Zhengzhou (within the red circle). The second covers Hong Kong and its surroundings (blue circle)[3]. The third comprises a vast but sparsely populated area in Central Mongolia (purple circle). Based on rainfall climatology computed with 24-hourly ERA5-ecPoint rainfall estimates between 1991 and 2020, the 1-year return period corresponds to $\sim$120 mm/24h for Zhengzhou and Hong Kong, and 15 to 35 mm/24h over Central Mongolia. Most areas on the map (including Hong Kong) show probabilities of exceeding the 1-year return period below 5%, with local peaks between 20-40%. In contrast, the areas over Central Mongolia and Zhengzhou display probabilities exceeding 50% over widespread regions, respectively, up to 60 and 80%. The map showing rainfall probabilities of exceeding the 50-year return period (Figure 7.8c) confirms the potential for extreme rainfall around Henan's capital and Central Mongolia. The probabilities of exceeding the 50-year return period in Zhengzhou (corresponding to approximately 500 mm/24h) remain high at up to 20%. The probabilities in Central Mongolia of exceeding the 50-year return period (ranging between

[3]For a detailed description of the heavy rainfall over Hong Kong due to the tropical storm Cempaka, please refer to: https://www.weather.gov.hk/en/wx-info/pastwx/mws2021/mws202107.htm

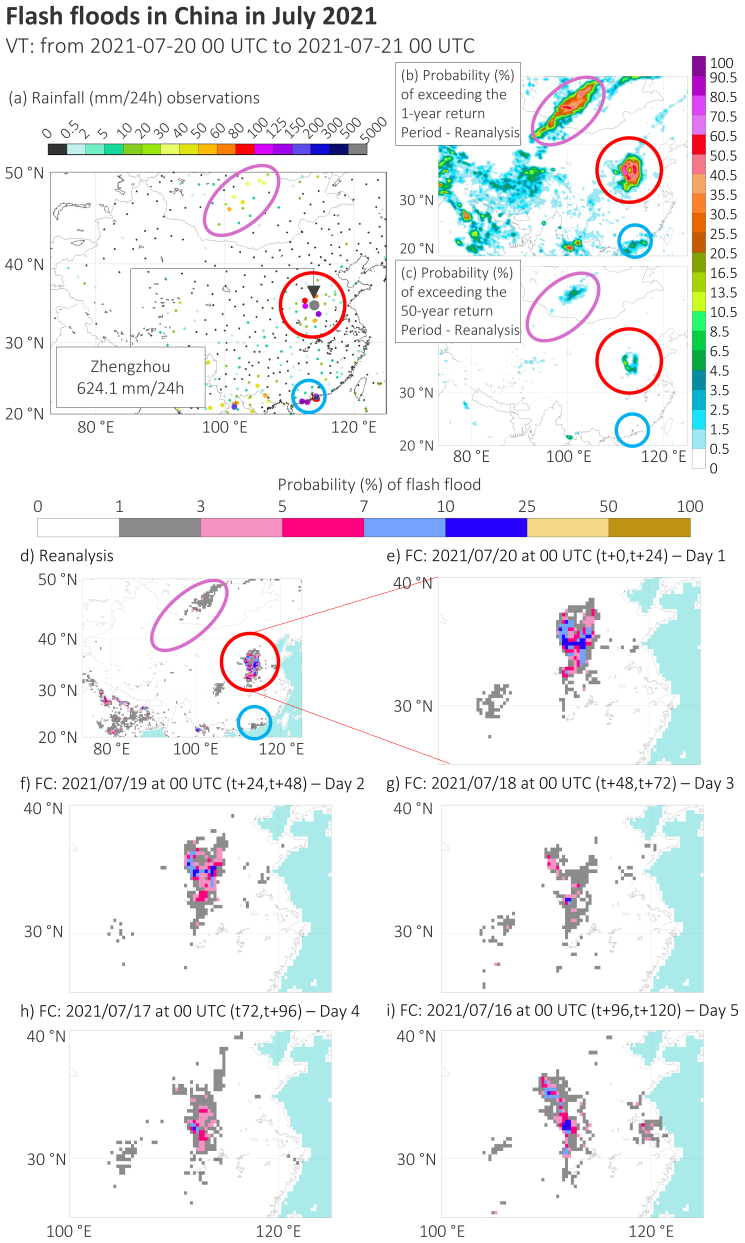

Figure 7.8: Flash floods in China in July 2021. The valid time (VT) for all plots is from the 20th of July 2021 at 00 UTC to the 21st of July 2021 at 00 UTC. Panel (a) shows rainfall observations (mm/24h). Panels (b) and (c) show the ERA5-ecPoint rainfall probabilities (%) of exceeding, respectively, the 1- and 50-year return period from reanalysis. Panel (d) show the probability (%) of flash flood computed with reanalysis data. The red circle highlights the high-to-extreme rainfall totals in the area surrounding and over Zhengzhou, the blue circle highlights the high rainfall totals over Hong Kong and surroundings, while the purple circle highlights an area with moderate rainfall over Central Mongolia. Panels (e) to (i) show the probability of flash flood computed from ERA5-ecPoint rainfall forecasts (FC), respectively, for day 1 - 2021/07/20 at 00 UTC (t+0, t+24) - day 2 - 2021/07/19 at 00 UTC (t+24, t+48) - day 3 - 2021/07/18 at 00 UTC (t+48, t+72) - day 4 - 2021/07/17 at 00 UTC (t+72, t+96) - and day 5 - 2021/07/16 at 00 UTC (t+96, t+120). These plots focus on the area surrounding and over Zhengzhou (within the red circle).

60 and 100 mm/24h) do not exceed 6%. The probabilities over Hong Kong of exceeding the 50-year return period (∼500 mm/24h) are 0%. ERA5-ecPoint rainfall forecasts provide a good signal for heavy rainfall over Zhengzhou and Central Mongolia up to day 5, while the signal around Hong Kong appears only from day 3 forecasts (not shown).

Flash floods in China in July 2021: data-driven, hydro-meteorological predictions of areas at risk of flash flood

The data-driven model, incorporating hydro-meteorological features, presents a different picture of areas at highest risk of flash floods. The data-driven output computed with reanalysis data (Figure 7.8d) reveals that only areas around Zhengzhou (red circle) show probabilities up to 10% of flash flood risk. These probabilities peak at 25% over the city itself. These high probabilities are due to high antecedent soil moisture, close to 100% due to persistent rainfall over previous days in the area, and the high probabilities of exceeding the 50-year return period (as shown by the shap values for the specific case, not shown). In contrast, the probabilities of flash floods over Central Mongolia and Hong Kong do not exceed 3%. These smaller probabilities are due to the low probabilities of exceeding the 1-year return period over Hong Kong, and the low probabilities of exceeding the 50-year return period over Central Mongolia (as seen on the shap values for the specific case, not shown). The data-driven output run with forecasts for the flash flood events around Zhengzhou demonstrates good predictive capability. Day 1 (Figure 7.8e) and day 2 (Figure 7.8f) forecasts maintain high probabilities up to 25% over the correct location. Earlier forecasts from days 3 to 5 (Figures 7.8g-i) also show a good signal for potential extreme flash flooding because the probability distributions over the general area remain similar to later forecasts. However, the locations with the highest probabilities are shifted south of Zhengzhou (due to the shift in the rainfall's peak location), suggesting that the less populated areas of Southern Henan may be those potentially affected by extreme rainfall and flash flooding.

7.7 Discussions

TA3 emerges as the most promising approach for developing predictions of areas at risk of flash floods from geographically-limited training data, especially when carefully selecting training regions that maximise hydro-climatological diversity and observational density.

Amongst all evaluated strategies, TA3 emerges as the most promising approach for developing predictions of areas at risk of flash floods from geographically-limited training data. Models exposed exclusively to data-rich regions during training have been shown here to extrapolate forecast probabilities in ungauged areas that resemble those as if the model had been trained with data over the whole domain. This is further supported by the probabilities computed with TA3-2 being very close to the baseline

probabilities (i.e., those computed with the full training dataset). Despite its reduced training data, the geographical relevance and relatively high-density observations in TA3-2 suggest that regional representation in training data proves more critical than absolute data quantity for capturing flash flood occurrence. Moreover, the specific data-driven model used in this chapter (XGBoost implementation for gradient boosting) has proved to learn general flash-flood-triggering hydro-meteorological patterns from reduced, but well-representative training datasets and preserve the integrity of the learned probability distributions, enabling reliable predictions when encountering similar hydro-meteorological conditions in previously unseen geographical domains. This finding aligns with evidence presented in Kratzert et al. (2024), who demonstrated similar transferability of learned hydro-meteorological relationships for riverine flood prediction, suggesting that data-driven approaches can capture fundamental hydrological processes that transcend specific geographical boundaries. However, not all regions with observations may perform equally. The higher performance of TA3-2 (training dataset over the *east* side of the CONUS) over TA3-1 (training dataset over the *west* side) demonstrates the critical importance of selecting regions that maximise hydro-climatological diversity and observational density. By utilising the eastern CONUS — a region characterised by diverse precipitation regimes, varied topography, and relatively high flash flood frequency — the model learned generalisable patterns that were successfully transferred to predict areas at risk of flash floods in unseen regions (in this case, the Western side of the CONUS).

The alternative training approaches (TA1 and TA2) reveal fundamental limitations that render them unsuitable for the global extension of regionally-trained models.

TA1 showed that full-domain but sparse coverage produces models with progressively degraded performance (both in reliability and discrimination ability) as the observational density decreases. The data-driven model became severely underperforming with large reductions in observational density (90% in TA1-3). In this case, the data-driven model trained with such a severe lack of observations cannot learn general hydro-meteorological patterns that may trigger flash floods. This is exemplified by the ROC curve built with operationally-relevant probability thresholds (not lower than 1%) that mostly overlap with the diagonal, which is typical of a model with no discrimination ability. The model's reliability is also very poor, showing a systematic shift towards severe under-prediction seen in the reliability diagram and in the frequency bias. The model's systematic failure in predicting areas at risk of flash floods appears even more evident by examining the precision-recall curves. When the curve displays only a short segment before terminating at low recall values, this indicates that the model has

learned to make positive predictions for only a minuscule fraction of the evaluation data, essentially defaulting to negative predictions beyond a very conservative probability threshold. This behaviour also manifests a paradoxical relationship between training data availability and initial precision: models trained with fewer positive reports often achieve higher precision at very low recall values compared to those trained on more comprehensive datasets. This counterintuitive result arises because extreme data scarcity forces data-driven models to become hyper-conservative, triggering positive predictions only under conditions that almost exactly match their limited training examples. Whilst this extreme selectivity yields high precision for the few predictions made, it comes at the cost of missing the vast majority of actual positive cases. In contrast, models trained with more diverse positive examples learn richer representations that enable generalisation across varied flash flood contexts. Although this broader learning results in lower precision at the beginning of the precision-recall curve because the model attempts to capture a wider range of conditions (also at the cost of including more false alarms), these models ultimately provide operationally viable performance by maintaining reasonable precision across meaningful recall levels. For flash flood prediction systems, where the failure to detect events has potentially fatal consequences, the ability to identify hazards across their full diversity of occurrence patterns far outweighs the superficial advantage of perfect precision limited to an insignificant fraction of events, underscoring the critical importance of guaranteeing exposure to sufficient flash flood occurrences to the model. Similar results, therefore, would be obtained if we were to train a data-driven model with global impact reports from databases such as EM-DAT. TA2 shows contrasting performances. TA2-2, despite being exposed during training to western regions appearing to have zero flash flood occurrences, successfully generalises learned patterns to these ostensibly flood-free areas, achieving performance metrics approaching those of the baseline model. This success stems from the eastern CONUS encompassing diverse hydro-climatic regions — from humid subtropical to continental climates — and experiencing relatively high flash flood frequency, providing sufficiently rich training signals to capture the fundamental relationships between atmospheric forcing, land surface characteristics, and flash flood occurrence. The model's ability to extrapolate these learned patterns to the semi-arid western regions, despite never observing positive examples there during training, suggests that the diversity and density of eastern observations can compensate for the artificial absence of western reports. Crucially, this compensatory capability proves highly sensitive to training data quality: when the same approach is

applied using western data (TA2-1), the limited hydro-climatic diversity and lower flash flood frequency causing less number of reports in the Western side of the CONUS fail to provide adequate training signals, resulting in an extremely poor performance over the unseen areas. This is also confirmed by the case study presented in the results section. The case study shows that this training approach reduces the probabilities of having a flash flood event close to 0 over the entire CONUS domain.

Objectively verifying the global data-driven predictions remains challenging due to the lack of observations

The absence of comprehensive global impact observations prevents rigorous global objective verification of the forecasts. Global disaster databases such as EM-DAT and DesInventar, though invaluable for large-scale impact assessment, exhibit severe reporting biases, typically capturing only events exceeding mortality or economic thresholds and systematically under-representing more minor flash floods that can also cause substantial local impacts. There also exist spatial biases, with reporting quality correlating strongly with institutional capacity, media coverage, and proximity to urban centres, precisely inverse to the distribution of vulnerability. This verification gap forces reliance on case study analyses, as demonstrated in this research, which, whilst providing valuable insights into model behaviour for specific high-impact events, cannot establish robust statistics about forecast performance critical to bridge the gap between methodological capability and operational confidence. Meteorological services must issue warnings based on models whose performance remains statistically unquantified in their regions, whilst international donors and governments must allocate resources to prediction systems whose local accuracy cannot be systematically demonstrated. Future solutions may emerge through adopting less standard observations to bulk current databases or create ones that take into account more minor local events but potentially equally impactful as those already included in the most commonly used impact databases. Satellite-detected flooding could be leveraged as a proxy for ground impacts. Crowdsourced observations from citizen scientists and social media streams could offer real-time flood impact information that could validate and calibrate predictions, yet extracting reliable information from noisy, biased social data requires advanced natural language processing and verification protocols. The integration of Internet of Things (IoT) devices, including low-cost weather stations and water level sensors, promises to dramatically expand observational coverage, though sustainability concerns regarding maintenance, calibration, and data transmission in resource-limited settings must be addressed.

Operational deployment may also be a challenge

Whilst this research establishes the methodological foundation for global prediction of areas at risk of flash flood, the transition from research demonstration to operational deployment presents challenges that merit careful consideration. Training a model operating at the spatial resolutions preferred in flash flood prediction (typically under 10 km) across global domains would require substantial computational resources, particularly if ensemble forecasts are developed to quantify uncertainty. Whilst at inference time, data-driven model are extremely cheap to run (even at high resolutions), training such a model might require considerable technological resources available only in a handful of centres such as ECMWF which already develops GloFAS. Model updating strategies present another layer of complexity—the continuous integration of new flash flood observations could improve prediction accuracy, but requires automated quality control procedures and regular retraining cycles (perhaps, as often as impact reports are updated, 4 to 6 times a year for NOAA's Storm Event Database, perhaps less frequent - once or twice a year for other databases) that demand both computational resources and human oversight.

Extension to multi-hazard early warning frameworks

The transfer learning framework demonstrated in this chapter for flash floods addresses a fundamental challenge shared across numerous hydrometeorological hazards, i.e., the existence of high-quality regional datasets alongside the absence of comprehensive global observations. Lightning strikes, for instance, benefit from dense ground-based detection networks in North America (National Lightning Detection Network[4]) and Europe (EUCLID[5]), yet vast regions across Africa, Asia, and South America remain unmonitored despite experiencing intense thunderstorm activity. Similarly, landslide inventories such as the USGS Landslide Inventory[6] in the US or the Italian IFFI database[7] offer detailed historical records within specific regions, whilst global databases remain sparse and inconsistent. The methodological framework established here — training models on hydrometeorologically diverse regions with quality observations, and then applying them globally — could be directly adapted to other hazards such as those just mentioned. For lightning prediction, models trained on the North

[4]https://www.vaisala.com/en/products/national-lightning-detection-network-nldn

[5]https://www.euclid.org/

[6]https://www.usgs.gov/tools/us-landslide-inventory-and-susceptibility-map

[7]https://www.progettoiffi.isprambiente.it/?lang=en

American or European observational datasets could learn relationships between atmospheric conditions and lightning occurrence, subsequently providing life-saving predictions in regions where agricultural workers and school children face significant lightning mortality without warning systems. Landslide susceptibility models developed using comprehensive inventories from mountainous regions in Europe or Japan could be transferred to predict this hazard in data-scarce mountainous areas across the Andes or Himalayas.

Climate change adaptation and model robustness under non-stationarity

The demonstrated transferability of flash flood predictions across diverse hydro-climatic regions carries profound implications for climate change adaptation strategies. As global warming intensifies the hydrological cycle, regions historically experiencing infrequent flash flooding may transition into high-risk zones, whilst traditional flood-prone areas may face unprecedented extremes. The success of models trained on the climatologically diverse eastern CONUS in predicting flash floods in semi-arid western regions suggests that current data-driven approaches could provide anticipatory warning capabilities for communities entering unfamiliar climate regimes. However, this raises critical questions about model robustness under non-stationary conditions: how frequently must models be retrained to capture evolving precipitation-runoff relationships, and can transfer learning approaches trained on present-day extreme events adequately predict future extremes that may exceed historical analogues? The answer to these questions is tentatively positive, because even though specific verification for flash flood occurrence is needed, positive signals in this direction have already been presented in the literature (Bertola et al., 2023). Moreover, through the development of a historical dataset of flash flood occurrence by running the data-driven model on retrospective reanalysis data, it would be possible to create climatological studies to assess increased/decreased frequency of flash floods in a particular region, or better understand the hydro-meteorological patterns that generate flash flood occurrence.

Spain and China case studies validate TA3-2 global transferability; physical mechanisms learnt over CONUS generalise to unseen regions

The case studies from Spain (October 2024) and China (July 2021) empirically validate how successfully the training approach TA3-2 can apply the general patterns learnt from geographically-limited, but representative, training data to predict areas at risk of flash floods in unseen regions with distinct hydro-climatic characteristics. This is exemplified by the use of antecedent soil moisture conditions and probability of exceeding extreme rainfall totals (i.e., 50-year return period) to modulate the final risk of flash floods in different regions over China and Mongolia. Or by considering the

risk of high rainfall totals in nearby areas to those of interest, to modulate the risk of flash floods suggested by only rainfall-based predictions. These are the same characteristics seen in predictions done over the CONUS with the data-driven model trained with data for that domain. These successful applications, achieved without any exposure to European or Asian flash flood impact reports during training, substantiate the research's central finding that carefully selected regional training data capturing sufficient hydro-climatological diversity can enable reliable global flash flood predictions. Similar results have been obtained for riverine floods (Kratzert et al., 2024).

7.8 Conclusions

This chapter has systematically investigated the critical trade-offs between spatial coverage and observational density in developing transferable flash flood prediction models, addressing Research Question 3 of this thesis. Through comprehensive sensitivity analysis across three distinct training approaches, we have demonstrated that the aspiration to achieve global flash flood prediction coverage is methodologically feasible despite the severe paucity of observations in most regions worldwide.

Training on diverse, data-rich regions (TA3) is optimal; sparse sampling and artificial absence corrupt model learning

The empirical evidence establishes that training over a domain with good observational data emerges (TA3) as the optimal strategy for extending regional models to global applications. This approach, whereby models train exclusively on data-rich regions before deployment to unseen areas, successfully preserves the integrity of learned hydro-meteorological relationships. The striking performance disparity between eastern-trained and western-trained models underscores, however, a fundamental principle: successful transfer learning requires training regions that encompass large enough hydro-climatological diversity and fairly high observation density. The eastern CONUS, with its varied precipitation regimes spanning humid subtropical to continental climates, provides the requisite diversity for models to learn generalisable flash flood generation mechanisms. Conversely, the systematic failures of alternative approaches illuminate critical constraints on global model development. Random spatial sampling (TA1) demonstrates that maintaining geographical coverage whilst reducing observation density produces catastrophic performance degradation, with models becoming hyper-conservative and operationally ineffective when positive examples become too sparse. The regionally-constrained training with global visibility (TA2) reveals an even more fundamental limitation:

exposing models to regions with artificially absent flash flood observations may also severely corrupt the learning process.

Remaining challenges: global verification gaps, multi-region data aggregation, non-standard observations, and extension to other hazards

Significant challenges remain before operational implementation can bridge the gap between methodological capability and life-saving warnings. For example, the absence of comprehensive global verification data prevents rigorous assessment of model performance precisely where it matters most, i.e., in vulnerable, data-scarce regions. In this chapter, impact reports only from the CONUS were considered; however, it is foreseen that strategic aggregation of data from diverse, well-monitored regions (such as Western Europe or selected countries such as Japan, Ecuador, or Brazil) could further improve the predictions computed for this thesis. It is also anticipated that innovative solutions, such as the use of other non-standard observations (satellite-derived inundation maps, crowdsourcing observations, and social media posts), will be applied to bulk current observational datasets, enabling objective validation analysis of global predictions. Moreover, it is also envisioned that the knowledge developed in this chapter about applying regionally-trained data-driven models to global applications might be extended to also low-probability, high-impact hazards such as landslides and lightning. Finally, as climate change drives communities into unfamiliar hazard regimes, the ability to transfer knowledge from observed to unobserved regions becomes increasingly critical for adaptation and resilience.

CHAPTER 8

GENERAL DISCUSSIONS

Escalating flash flood impacts and climate-change-induced intensification of flash flood risk demand urgent advances in forecasting capabilities

Flash floods represent the deadliest and most devastating form of natural hazard worldwide, causing over 5,000 fatalities annually and accounting for approximately 85% of global flood incidents. Recent catastrophic events, such as those including (but not limited to) the flash floods in the USA, Spain, Libya, Germany/Belgium, Central Asia, and Brazil, have claimed thousands of lives, caused countless injuries, and resulted in billions of dollars in economic losses. As climate change continues to intensify the frequency and severity of extreme rainfall - including in historically low-risk regions - the development of accurate, timely, and globally accessible flash flood predictions has become a critical priority for disaster risk reduction. WMO has identified flash floods as one of its top priority natural hazards. Meanwhile, the UN's 'Early Warnings for All' initiative, launched in 2022 with the ambitious goal of protecting every person on Earth with early warning systems by 2027, places flash floods at the forefront of its agenda.

Thesis aim: addressing technical barriers to medium-range predictions of areas at risk of flash floods over a continuous global domain to provide a viable pathway for protecting vulnerable communities worldwide

Significant technical and methodological obstacles persist in developing medium-range flash flood forecasts over continuous global domains. The overall aim of this thesis was to address the critical gap in global flash flood early warning capabilities and to demonstrate how data-driven approaches, combined with global numerical weather prediction models, can provide a viable pathway for protecting vulnerable communities worldwide, particularly those in data-scarce regions where traditional forecasting approaches face significant limitations. These three interconnected research objectives have been addressed in the main analysis chapters of this thesis:

Chapter 5: depart from the traditional rainfall-to-rainfall verification approach and adopt, instead, a *flash-flood-focused verification framework* that directly compares rainfall forecasts with flash flood impact reports to answer research question n.1 (RQ1) ”Can post-processed global NWP rainfall forecasts successfully identify areas at risk of flash floods up to medium-range lead times?”

Chapter 6: develop data-driven models that integrate hydro-meteorological variables from global reanalysis and global medium-range NWP forecasts, and flash flood impact reports to predict areas at risk of flash floods, from short (i.e., day 1) to medium-range lead times (i.e., day 5), to answer research question n.2 (RQ2) ”Are medium-range data-driven hydro-meteorological predictions of areas at risk of flash floods feasible with global reanalysis, forecasts, and impact flash flood reports?”

Chapter 7: assess how varying spatial coverage and data density scenarios may influence training strategies when creating global predictions with regionally-trained data-driven models, to answer research question n.3 (RQ3) ”How does coverage-density trade-off influence training data strategies to develop predictions of areas at risk of flash floods over a continuous global domain?”

This chapter discusses the outcomes, implications, and limitations of the research from each of these three main analysis chapters, synthesises the broader contributions to flash flood prediction science, and presents recommendations for advancing global early warning capabilities against flash floods that could ultimately save thousands of lives every year and support UN’s vision of global early warning coverage - for flash floods.

8.1 Development of a flash-flood-focused verification framework for predictions of areas at risk of flash flood against flash flood impact reports

Contribution to knowledge no. 1: creation of a flash-flood-focused verification framework and a benchmark verification dataset for data-driven predictions based on hydro-meteorological parameters

By addressing RQ1, this thesis introduces a *flash-flood-focused verification framework*, a methodology that directly compares global NWP

rainfall predictions with observed flash flood events. Departing from conventional rainfall-to-rainfall verification paradigms, this contribution addresses a methodological gap in the literature given by the implicit assumption that improved rainfall forecasts necessarily enhance flash flood predictions. By implementing this direct comparison between global NWP rainfall predictions and observed flash flood events, we provide empirical evidence of whether advances in numerical weather prediction modelling translate to enhanced flash flood predictability up to medium-range timescales. The investigation employed the CONUS as the primary experimental domain, leveraging the exceptional spatio-temporal resolution and completeness of NOAA's Storm Event Database for objective verification. Moreover, the outcomes of this activity provided a performance benchmark for the short- and medium-range data-driven flash flood predictions developed at the later stages of this research.

8.1.1 Key insights and contributions

ERA5-ecPoint identifies flash flood risk with meaningful skill to day 5, challenging the assumption that high-resolution forecasts are required

The most significant finding from this research is that global NWP models, specifically ERA5, when post-processed through the ecPoint technique, can indeed identify areas at risk of flash floods with meaningful skill up to medium-range (day 5) lead times. The values of AUC-ROC consistently above 0.6 across all lead times establish that useful predictive information persists throughout the medium-range forecast horizon. Hence, this finding challenges the prevailing assumption that flash flood prediction must require high-resolution forecasts. Similar results were also obtained by Pillosu et al. (2024) when ECMWF's ENS rainfall forecasts and their post-processed version with ecPoint were evaluated against flash flood reports in Ecuador.

Optimal reliability at 20–50-year return periods; ambiguity between genuine skill and reporting bias remains unresolved

The identification of optimal return period thresholds also represents a crucial operational contribution. The analysis reveals that when extreme rainfall events (inferred using the ERA5-ecPoint 99th percentile) exceed the 20- and 50-year return periods, we achieve frequency bias values, versus the flash flood reports, of between 2 and 3, so much closer to the optimal value of unity than is the case for shorter return periods. This convergence towards optimal reliability at extreme thresholds provides two potential interpretations that merit consideration. First, the rainfall-based predictions may genuinely perform best when focused on the most extreme events, as these create sufficiently severe hydrological conditions to overwhelm local modulating factors such as soil moisture or drainage capacity. Second, the Storm Event Database may predominantly capture only the most severe

flash flood events, with routine or moderate events systematically underreported. The inability to definitively distinguish between these interpretations highlights a fundamental challenge in verification against impact databases.

Verification framework is transferable to other rainfall predictions and hazards with minimal adaptation

The verification framework developed in Chapter 5 itself constitutes a transferable methodological contribution that extends beyond the specific findings for the considered global rainfall forecasts (ERA5-ecPoint) and the considered hazard (flash floods). By establishing standardised procedures for transforming point-based impact reports into gridded observational fields, defining verifying thresholds from climatological data as done in this thesis or from short-range forecasts as shown by Pillosu et al. (2024)), and computing both reliability and discrimination metrics, this research provides a blueprint for evaluating any rainfall predictions (see examples from (Pillosu et al., 2024)) and other hazards with minimal adaptations to the methodology (see examples for severe weather in Robbins and Titley (2018) and Tsonevsky et al. (2018) and Cavaiola et al. (2024) for lightning). Possible modifications to the methodology proposed here might include ways to account for location and time uncertainties in the impact reports if forecasts with much higher spatial and temporal resolution are considered. In this specific development, such uncertainties did not play a crucial role as the forecasts were provided over ERA5's low spatial resolution grid ($\sim$31 km) and over 24-hourly periods, which can counterbalance the inherent uncertainties in reporting.

8.1.2 Limitations

Verification limited to CONUS; return period thresholds may not generalise to other regions

The verification was conducted exclusively over the CONUS, leveraging the exceptional quality and spatial coverage of NOAA's Storm Event Database. While this does not limit the potential transferability of both the verification framework and the established performance metrics, the return period thresholds found in this thesis might not hold in other parts of the world.

Impact reporting biases inflate false alarm rates, particularly in rural and disadvantaged areas

Some issues can also be anticipated over the CONUS. Despite the apparent completeness of the Storm Event Database, inherent biases in impact reporting — including population density effects, diurnal reporting variations, and socio-economic disparities in hazard documentation (Marjerison et al., 2016) — introduce systematic uncertainties that the developed verification framework cannot fully quantify. The fundamental assumption that the absence of a report equates to the absence of an event inevitably

inflates false alarm rates, as locations experiencing flash floods without documentation are classified as false positives. This effect compounds in rural or economically disadvantaged areas where reporting infrastructure is limited, creating a spatially heterogeneous bias field that may correlate with the very vulnerabilities the warning system aims to address. The resulting frequency bias values, while useful for relative comparison across thresholds and lead times, inevitably present an imperfect picture of true forecast performance.

Rainfall-only approach neglects hydrological modulation; incorporation of hydrological parameters is needed

The rainfall-only approach, whilst demonstrating good predictive skill, fundamentally neglects the hydrological processes that modulate rainfall-runoff transformation. Antecedent soil moisture conditions, land surface characteristics, urbanisation extent, and drainage network properties all influence whether a given rainfall event triggers flash flooding, yet these factors remain unaccounted for in this chapter. By neglecting these factors, the rainfall-only approach may miss critical flash flood events in pre-saturated catchments while generating false alarms in areas with high infiltration capacity or robust drainage systems. Hence, the scientific and operational community should strive to develop global forecasts that also incorporate hydrological parameters to better identify areas at risk of flash floods.

8.1.3 Future research directions

Advancing the flash-flood-focused verification framework requires coordinated efforts across multiple research fronts.

Priority: apply verification framework to other data-rich regions to test threshold universality

An immediate priority may involve applying the verification framework in other data-rich regions. Verification studies using the European Severe Weather Database over Europe, the Japanese Meteorological Agency records, or the Australian Bureau of Meteorology flood databases would test the universality of the established thresholds for flash-flood- triggering return periods. These studies should explicitly examine how verification metrics vary with different impact reporting systems, hydro-climatic regions, and societal factors affecting report generation.

Separating verification by convective regime would yield more nuanced predictability assessments

Another aspect that could be considered is separating verification statistics for large-scale and convective systems. Their predictability is very different (Pillosu et al., 2024), and this might impact the outcomes of the verification. The findings of Pillosu et al. (2024) demonstrate significant performance disparities in flash flood prediction across different convective regimes. Their verification statistics reveal that regions dominated by

large-scale convective systems — which tend to be represented in global NWP model outputs as large-scale or mostly large-scale rainfall — exhibit markedly better forecast skill in raw model output compared to regions impacted primarily by small-scale convective systems. Conversely, post-processed forecasts with the ecPoint methodology showed their true benefit in the second case, by enhancing the identification of areas at risk of flash floods missed by the raw forecasts. These contrasting outcomes suggest that incorporating in the verification framework rainfall-related discriminators (such as whether the flash flood event was generated by a large- or small-scale system) could provide more nuanced insights into the performance and predictability of global NWP rainfall forecasts for the identification of areas at risk of flash flood.

ML-based post-processing and impact database bias correction could enhance forecast quality and verification accuracy

Alternative post-processing techniques warrant investigation, particularly machine learning-based downscaling methods that could reduce computational demands whilst potentially enhancing forecast quality. For example, deep learning architectures trained on the relationship between coarse-resolution NWP outputs and high-resolution precipitation observations might in time be able to replace the weather regime matching of ecPoint, potentially at lower cost. Finally, developing bias correction methodologies for impact databases—perhaps through integration of satellite-based flood detection, social media mining, and population density weighting—could address the systematic underreporting that affects verification metrics, providing more accurate assessments of true forecast performance.

8.2 Development of data-driven hydro-meteorological predictions of areas at risk of flash flood

Contribution to knowledge no. 2: data-driven hydro-meteorological flash flood prediction handling severe class imbalance

The advancement from rainfall-only predictions to an integrated hydro-meteorological data-driven approach represents a critical evolution in flash flood forecasting methodology. Chapter 6 addressed Research Question 2 (RQ2) by demonstrating the feasibility of developing data-driven predictions that successfully handle severe class imbalance in the observation dataset whilst incorporating multiple environmental variables to enhance predictive capability. The systematic evaluation of six machine learning architectures, combined with comprehensive hyperparameter optimisation and feature

importance analysis, establishes both the potential and the constraints of data-driven approaches for operational flash flood prediction at a regional scale.

8.2.1 Key insights and contributions

Machine learning extracts meaningful flash flood signals from datasets with only 0.27% positive cases

The successful development of a data-driven model using a severely imbalanced observational dataset - with flash flood events comprising merely 0.27% of all cases - demonstrates that carefully configured machine learning algorithms can extract meaningful predictive signals for flash flood detection from sparse observational datasets, up to medium-range lead times. A few considerations emerge, however, from this achievement.

Simpler architectures suffice for large-domain binary prediction; XGBoost optimal for performance and efficiency

The comprehensive evaluation of multiple architectures (decision-tree-based, such as random forest and gradient boosting, and feed-forward neural networks) reveals that gradient boosting implementations and shallow neural networks were able to learn functional patterns in the data to predict areas at risk of flash floods with good discrimination ability and reliability up to medium-range lead times. The analysis of the hyperparameter tuning suggests that these simpler architectures were already able to extract the patterns in the data that were needed for the successful prediction of areas at risk of flash floods over a large domain (i.e. the CONUS) and up to medium-range lead times. This is particularly exemplified in the neural networks. Whilst during optimisation, the neural network was allowed to have multiple (>2) hidden layers, the optimisation suggested that one hidden layer was enough to generalise from the data at hand and identify areas at risk of flash floods, even from medium-range forecasts. This result appears to be in striking contrast - even if different fields than flash flood forecasting - with the current literature on deep learning for flash flood prediction, suggesting that deep neural networks outperform simpler implementations, such as random forests, gradient boosting and shallow neural networks (Roy et al., 2020; Kumari and Toshniwal, 2021). These studies differ in nature from this research. They aim to predict river discharge for specific catchments to generate high-resolution (< 1 km) inundation maps, thereby identifying potential flooded areas with high precision at very short lead times. The target of this thesis is simpler - it predicts the binary condition of yes or no flash flood occurrence without claiming to be more precise in terms of magnitude and timing of the event. Perhaps, it is for this reason that simpler models can satisfy the basic needs of operational flash flood prediction over large domains (Zanchetta and Coulibaly, 2020) - the prediction target of this

thesis - and meanwhile extend the time horizon of the forecasts from day 1 to day 5. In particular, XGBoost achieved the best balance between predictive performance and computational efficiency, with training times in the order of minutes; neural network training took hours, yet the resulting predictions were no better. In operational hydrology, such a short training time is highly appealing for re-training as more observations become available, or more features may need to be tested.

Weighted loss functions boost hit rates but degrade reliability; conservative approach preferred for operational credibility

General and imbalanced-data-specific loss functions were considered to develop the data-driven models presented in Chapter 6. Together with the manipulation of the training dataset to reduce the imbalance between the binary events, the use of loss functions specific for imbalanced datasets is considered standard practice in machine learning to help the data-driven model learning from the features to predict the target variable (Altalhan et al., 2025). The first approach - manipulating the training dataset by creating synthetic yes-events or removing possibly very informative non-events - was not considered in this thesis to avoid adding more uncertainties than those already there. However the second approach - considering loss functions specific for imbalanced datasets (i.e. weighted loss functions) - was considered alongside training the model with no specific treatment of the imbalanced characteristic. This comparison highlighted that models employing weighted loss functions successfully enhance detection of positive events, achieving hit rates approaching 90% for gradient boosting implementations. However, this enhanced sensitivity also resulted in false alarm rates exceeding 50%, representing a twenty-fold increase compared to using a loss function designed for balanced data. For flash flood prediction, where public trust and warning fatigue represent critical operational constraints, the conservative approach using balanced loss functions emerges as preferable, despite lower hit rates (20-30%). This consideration challenges the assumption that maximising hit rates should be the primary objective in rare event prediction, highlighting instead the importance of maintaining forecast credibility through balanced performance. Moreover, identification of the areas at most risk of flash flooding was made difficult by the increased noise created by the models trained with weighted loss functions. Whilst, from a verification standpoint, this was not problematic because it did not always show a worsening of the scores, in practice, such forecasts would hold lower value for practitioners (e.g. forecasters) because areas at risk of flash floods would not be highlighted as precisely.

Hydrological features provide 10–35% contribution to predictions; all selected features retained during optimisation

The integration of hydro-meteorological features beyond precipitation represents a fundamental advance over the rainfall-only baseline established in Chapter 5, and typically used in operational hydrology when identifying areas at risk of flash flood over large domains. Whilst the SHAP analysis confirms that rainfall probability for the 1-year return period dominates model decisions with an average 80% contribution, the secondary features also hold high importance with contributions of 10-35% to the mean absolute SHAP values, and provide critical modulation of flash flood risk as seen by the dependency maps. The importance of considering hydrological parameters for predicting areas at risk of flash floods is evident in the *low importance* given to hyperparameters that specify the fraction of features to be dropped while training gradient boosting (hyperparameter ”colsample_bytree”) and neural networks (hyperparameter ”dropout”). Specifically, values of these hyperparameters mostly remained close to 1 during the multiple optimisations, indicating that the selected features were all important, although at different levels as shown by the SHAP analysis.

Skill degradation with lead time enables tiered operational protocols from protective actions to preparedness

The temporal analysis of forecast skill degradation with lead time establishes practical boundaries for operational utility. The minimal performance reduction from reanalysis to day 1 forecasts demonstrates that short-range (up to t+24) predictions retain good accuracy for high-confidence protective actions (even if this is partly because ”reanalysis” rainfall values have to be created by summing segments of different short-range forecasts, up to 12h leads). Then the gradual degradation from day 3 to day 5 forecasts, where discrimination ability remains above 0.7, indicates that medium-range forecasts retain value for preparedness activities despite reduced precision and reliability. This temporal stratification enables differentiated response protocols where immediate evacuations or other high-cost actions such as resource mobilisation might be triggered by day 1 (or day 2) forecasts, whilst day 3-5 predictions may support low-cost, no-regret actions, such as public awareness campaigns.

8.2.2 Limitations

ERA5’s 31 km resolution cannot capture sub-grid hydrological variability critical for flash flood generation

The spatial resolution constraint imposed by ERA5’s 31-km grid boxes represents a fundamental limitation for capturing the localised processes that generate flash floods. While the ecPoint post-processing enhances the representation of rainfall’s sub-grid variability, and whilst the static variable representing orographic slope has a sub-grid component, the underlying hydrological variables are ostensibly at coarse resolution. Small-scale

variations in soil moisture, land cover, and drainage patterns that critically influence flash flood generation occur at much higher resolution scales than ERA5's resolution. This resolution mismatch between the scale of forecast information and the scale of flash flood processes inevitably limits predictive accuracy.

Single-realisation probability estimates lack rainfall uncertainty; operational ensemble inputs would likely improve skill

The fact that a single probability value is created here, for nominal operational decision making, using just one forecast realisation, could be viewed as a weakness. Although the individual algorithms employed, such as random forest and gradient boosting, are ensemble methods internally, the absence of the use of forecast model ensembles as inputs probably degrades performance. Note that operational ecPoint forecasts, although not employed in this thesis, would be available for any real-time forecast system, and these do amalgamate together post-processed forecasts from all 51 operational ECMWF ensemble members. In that sense, the skill levels quoted here probably represent a lower bound on what could be operationally achieved. Indeed, rather than being provided with a "deterministic" estimate of flash flood probability, emergency managers would be better served by having, within that value, an inbuilt representation of rainfall-prediction-related confidence, to provide better guidance on an appropriate response. The absence of this rainfall uncertainty information particularly affects medium-range forecasts and predictions over less inherently predictive systems (e.g., small-scale convective systems, as opposed to flash floods from large-scale, organised convective systems, where ensemble member spread might typically be less).

SHAP reveals counter-intuitive vegetation and orography relationships, likely reflecting seasonal coincidence and reporting bias

Several counterintuitive relationships revealed through SHAP analysis raise concerns about the physical interpretability of model behaviour. The positive correlation between vegetation density and flash flood probability contradicts established hydrological understanding, where vegetation typically increases infiltration and reduces runoff. While this may reflect seasonal coincidence between maximum leaf area index and peak convective season, such spurious correlations risk producing the correct answers for the wrong reasons. Similarly, the indication that flatter terrain increases flash flood probability conflicts with physical hydrology but may reflect reporting biases where impacts concentrate in wide populated valleys and other relatively flat terrain, rather than steep headwaters where floods initiate. These interpretation challenges highlight the danger of purely data-driven approaches learning patterns in observational biases rather than underlying physical processes.

Storm Event Database biases may limit model applicability beyond the US socio-economic context

The exclusive reliance on the Storm Event Database for model training may introduce systematic biases in the data-driven outputs, limiting their applicability beyond the specific socio-economic context of the United States. Population density effects, whereby densely populated areas generate disproportionate reports relative to flood frequency, may become embedded in the model. Economic factors affecting reporting rates, infrastructure quality influencing impact severity, and cultural attitudes toward hazard documentation may also shape the observational dataset in ways that may not transfer to other regions.

Operational maintenance requires sustained computational and human resources, especially at higher resolutions

The computational requirements for maintaining currency in operational settings present practical challenges, especially if, in future applications, the forecasts' spatial resolution is higher (operational ecPoint forecasts currently act at 18km resolution, for example). While individual model training requires only 20 minutes (at 31 km spatial resolution), the need for periodic retraining as new events accumulate, combined with hyperparameter re-optimisation and performance validation, creates substantial computational demands. Moreover, the technical expertise required for model updates, regular verification of predictions, and physical interpretation of model outputs necessitates sustained investment in appropriate resources to provide forecasts with 24/7 operational assistance.

8.2.3 Future research directions

Ensemble approaches needed for robust uncertainty quantification in operational risk-based decision-making

Immediate priorities should focus on enhancing model sophistication through ensemble approaches that provide robust uncertainty quantification. Implementing bootstrap aggregation, model stacking, or Bayesian neural networks would generate prediction intervals essential for risk-based decision-making. The combination of multiple base learners through stacking could leverage the complementary strengths of different algorithms while providing natural uncertainty estimates through model disagreement. Deep ensemble methods, training multiple neural networks with different random initialisations, offer another pathway toward uncertainty-aware predictions suitable for operational deployment.

Higher-resolution NWP models approaching convection-permitting scales could better capture flash-flood-triggering processes

Integration with higher-resolution atmospheric models represents a critical pathway for improving predictive accuracy. As global NWP systems approach convection-permitting resolutions, e.g. with ECMWF's IFS at 9 km resolution and Destination Earth's 4 km implementation, the opportunity emerges to capture better the mesoscale processes generat-

ing intense rainfall. Coupling data-driven models with these enhanced hydro-meteorological predictions, while incorporating high-resolution static datasets for topography and land cover, could substantially improve the representation of flash-flood-triggering conditions.

Causal inference techniques could resolve spurious correlations and improve physical interpretability and generalisation

Addressing the challenge of spurious correlations between model features and their impact on the model output requires careful thought. The models tested in this thesis rely primarily on correlation learning, i.e., the ability to find patterns in data to make predictions. However, this capability can be limited in situations where a deeper understanding of the underlying causal relationships is required. This corresponds to our case, where it was found that more vegetated and flat areas increase the probabilities of having a flash flood event when physics tells us they should have the opposite effect. One approach to solving the problem could be to introduce additional types of observations, such as river discharge, to complement the information provided by impact reports and teach the model more diverse relations between the model features and the (varied) observations. Another approach could consider introducing causal inference techniques to distinguish genuine causal relationships from spurious correlations. The goal of causal inference in machine learning is to improve the accuracy and interpretability of models, fundamental prerequisites for implementation and adoption in operational contexts. It can also help to improve generalisation, which would help transfer the learnt patterns to unseen regions (to protect the most vulnerable communities in the Global South) or unseen hydro-meteorological conditions (due to climate change). Structural causal models, particularly directed acyclic graphs (DAGs), could explicitly encode known hydro-meteorological relationships, such as the protective effect of vegetation on runoff generation and the role of terrain slope in flow accumulation. Additionally, counterfactual reasoning frameworks could enable the model to answer critical ”what-if” questions, such as estimating how flood risk would change under different land cover scenarios while holding other factors constant. These causal inference approaches would complement the correlation-based learning by imposing physically-consistent constraints on the model’s decision-making process, ultimately producing predictions that align with both observed data and established hydrological understanding.

Transfer learning and multi-regional database integration needed for global applicability

Expansion beyond the CONUS domain necessitates strategic approaches to transfer learning and domain adaptation. Systematic evaluation of model performance in regions with similar hydro-climatic characteristics

but different reporting systems would better establish the extent of transferability. Fine-tuning pre-trained models with limited local observations could adapt learned patterns to regional specificities while preserving general hydro-meteorological relationships. The development of global training datasets through careful integration of multiple regional databases, accounting for reporting standard differences through appropriate normalisation, represents a longer-term goal for truly global applicability.

Non-standard data sources offer objective validation and improved observational density in data-sparse regions

The integration of alternative, non-standard data sources could address limitations in traditional impact databases while providing near-real-time validation capabilities. Satellite-based flood detection algorithms applied to imagery could provide objective flood extent information independent of human reporting. Social media analytics, particularly geolocated images and text descriptions of flooding, offer crowd-sourced validation data with unprecedented spatial and temporal coverage. Internet of Things sensors, including low-cost water level monitors and connected weather stations, promise to increase the density of observational networks in data-sparse regions.

8.3 Global implementation of regionally-trained data-driven models for flash flood prediction

Contribution to knowledge no. 3: transfer learning enables global flash flood prediction from regionally-trained models

The transition from regional to global flash flood prediction represents the culmination of this thesis's methodological development, addressing the fundamental challenge of extending data-driven approaches beyond data-rich domains. Chapter 7 systematically investigated Research Question 3 (RQ3) by examining how the trade-off between spatial coverage and observational density influences training strategies for developing predictions over a continuous global domain. Through a comprehensive sensitivity analysis of three distinct training approaches applied over the CONUS domain, this research contributes to the UN's "Early Warning for All" vision of universal early warning coverage. In this case, specifically for flash floods.

8.3.1 Key insights and contributions

Domain-exclusive training on data-rich eastern CONUS (TA3-2) optimal; validated in Spain and China without local training data

The most significant finding establishes that domain-exclusive training on carefully selected data-rich regions (TA3) emerges as the optimal strategy for global model deployment. The better performance of this

approach suggests that models trained exclusively on high-quality observations from limited geographical areas successfully preserve the integrity of learned hydro-meteorological relationships while maintaining predictive performance comparable to models trained on complete datasets. Global predictions were created adopting the TA3-2 approach. The XGBoost model (trained with a generic loss function for balanced datasets, and hyperparameters optimised by maximising the AUC-ROC metric) was trained only over the CONUS, with impact reports from the Storm Event Database over the CONUS. At the inference stage, the model was run with global data (from reanalysis and forecasts) to create global predictions. The empirical validation through case studies in Spain (2024) and China (2021) suggests that such a model could accurately predict flash flood risk in regions with fundamentally different hydro-climatic characteristics, without any exposure to local observations during training. The critical importance of training region selection becomes evident through the performance disparity between eastern-trained (TA3-2) and western-trained (TA3-1) models. Eastern CONUS, encompassing diverse precipitation regimes from humid subtropical to continental climates and experiencing relatively high flash flood frequency, provides sufficiently rich training signals to capture generalisable flash flood generation mechanisms. The model's successful extrapolation to semi-arid western regions and international domains validates the hypothesis that hydro-climatological diversity and observational density in training data matter more than absolute geographical coverage. This finding aligns with recent advances in large-sample hydrology (Kratzert et al., 2024), suggesting that data-driven approaches can capture fundamental physical processes that transcend specific geographical boundaries.

Sparse global databases (TA1) cause catastrophic degradation; regional absence of reports (TA2) introduces spurious spatial biases

The systematic evaluation of alternative training strategies provides crucial insights into the limitations of sparse global observations. TA1, whereby flash flood impact report density is reduced progressively to very low values over the whole domain, simulating conditions analogous to using global databases like EM-DAT, demonstrates catastrophic performance degradation as observational density decreases. Models trained with 90% data reduction become hyper-conservative, producing operationally useless predictions that default to zero or extremely low probability thresholds. This finding has profound implications for global flash flood prediction efforts, suggesting that current global impact databases lack sufficient density to support effective model training, even with state-of-the-art machine learning approaches. TA2, whereby a model is exposed to the whole domain of interest but certain parts never see flash floods, also shows poor overall

performance. A data-driven model might perform well over the regions where the data is available if the quality of the training data is good (TA2-2). If the quality is poor, a poor performance is also observed over the region from which the training data comes (TA2-1). However, in the former case, prediction quality deteriorates significantly over areas lacking flash flood reports. The model appears to learn an artificial distinction between regions, incorrectly inferring that the hydro-meteorological conditions triggering flash floods in the data-rich area may be less hazardous in the data-sparse areas. This spurious regional dependency emerges from the training data structure rather than genuine physical differences in flash flood susceptibility.

Transfer learning approach extensible to other hazards including landslides and lightning

The successful application of regionally-trained models over a continuous global domain does not stop at flash flood prediction. It instead offers the opportunity to maximise the value of regional observational datasets across a wide spectrum of hydro-meteorological hazards (e.g., landslides, lightning). For example, landslide susceptibility models developed using comprehensive inventories from Japan or Italy could potentially transfer predictive capability to mountainous regions across the Global South. Dense observational networks for lightning in North America and Europe could be leveraged to provide life-saving warnings in unmonitored regions.

8.3.2 Limitations

Global impact databases too biased for objective verification; case-study validation compelling but cannot establish systematic statistics

The absence of comprehensive global verification data represents the most critical constraint on establishing operational confidence in globally deployed models. It is well known that global impact databases exhibit severe reporting biases, capturing predominantly high-impact events while systematically underrepresenting smaller-scale flash floods that nonetheless can cause substantial local impacts (Panwar and Sen, 2020). This characteristic makes global impact databases unsuitable for providing robust statistics for a global model within an objective verification framework. The primary problem would be the inflated estimation of false alarms. They might be high due to a poor performance of the model itself, but more probably false alarms may appear extremely high due to the enormous number of missing reported events in the impact databases, causing the model to appear less performant than it would be if flash flood reports were included in global databases more consistently and with no biases. This is consistent with the verification results shown by Pillosu et al. (2024) over Ecuador. In that study, false alarms and frequency bias were extremely elevated, suggesting that the considered rainfall-based predictions of areas at risk of flash floods

were largely overestimating the observed events. Even though Pillosu et al. (2024) showed an overall good correspondence between regions with wet conditions in the forecasts and locations with flash flood occurrence in the regional impact database, the number of grid-boxes with reports was orders of magnitude less than those with a yes-event in the forecasts, causing the false alarms to be overwhelmingly high and producing a frequency bias in the thousands (where it should be around 1 to indicate well calibrated predictions). This situation would be identical if the global predictions were evaluated with global impact databases, but at a scale hundreds of times larger, making the predictions appear completely unusable. Thus, case-study-based subjective verification remains mostly the only practical pathway to assess model performance for globally deployed models. However, it is important to stress that, while case studies provide compelling evidence of successful predictions for specific cases (such as those presented in this thesis for Spain and China), they remain isolated examples that struggle to establish systematic performance statistics.

Transferability to truly data-scarce regions with unique hydro-climatic characteristics remains untested

The transferability assumptions underlying the global application of regionally-trained models may not hold uniformly across all hydro-climatic regimes. While the successful predictions in Spain and China demonstrate transferability from North American training data to European and Asian contexts, these remain regions with some structural similarities to the training domain. The model's performance in truly data-scarce regions – such as sub-Saharan Africa or small island developing states – remains untested. These regions often exhibit unique hydro-meteorological characteristics, infrastructure configurations, and vulnerability patterns that may not be adequately represented in the training data built from the Storm Event Database.

8.3.3 Future research directions

Strategic aggregation of diverse regional datasets could improve predictions for high-impact, low-probability events globally

Immediate research priorities should focus on strategic aggregation of high-quality regional datasets to create more representative global training samples. Combining observations from climatologically diverse regions such as Europe, Japan, Australia, and selected well-monitored areas in Latin America and Africa could provide training data that better represents global hydro-meteorological diversity. Moreover, such a dataset would likely include events that may lead to moderate-impact flash flood events in one region. Still, they might represent rarer or unseen conditions somewhere else, allowing us to provide better predictions for high-impact, low-

probability (or unseen) flash flood events. Such an approach recalls the *remote calibration* technique adopted in the ecPoint post-processing, which has been shown to be invaluable in the prediction of extremely rare or unseen rainfall events (Hewson and Pillosu, 2021). This perspective also aligns with the growing consensus in the literature to widen the spatial horizons of training datasets during model development (Kratzert et al., 2024) and our spatial perception of extreme event occurrence (Bertola et al., 2023) to issue more informed warnings for high-impact, low-probability (or unseen) events.

ML uncertainty quantification could guide targeted observational network expansion and synthetic data augmentation

The data-driven model deployed globally could also be used beyond predictive tasks. Strategic expansion of observational networks guided by machine learning uncertainty quantification represents an innovative approach already encouraged in the field of classical physics-based weather forecasts[1]. Such an approach could be used to improve our current global flash flood prediction capabilities. By analysing where trained models exhibit the overall poorest performance (for example, over Africa) or conditions where prediction uncertainty is at its highest (for example, when predicting localised flash floods due to small-scale convective systems over complex terrain), it would be possible to identify priority regions for new observation infrastructure deployment or real-time data collection from media. Thus, this targeted approach to network expansion could optimise limited resources by focusing investments where they provide maximum improvement in model performance. For example, experiments with generative AI techniques to synthetically augment the observational datasets could be used for this purpose. However, careful validation protocols must ensure that synthetic events remain physically plausible and do not introduce artificial biases that could compromise model reliability. The combination of uncertainty-guided observational network design and synthetic data generation could accelerate progress toward comprehensive global flash flood prediction coverage while making efficient use of limited resources for the deployment of new observational infrastructures.

[1]https://www.ecmwf.int/en/about/media-centre/news/2025/impact-experiments-support-initiative-more-weather-observations

CHAPTER 9

GENERAL CONCLUSIONS

Thesis motivation: flash floods cause >5,000 deaths annually; climate change intensifies the need for global early warning

Flash floods claim over 5,000 lives annually and represent the most devastating form of flood hazard worldwide, with their impacts transcending traditional divides between Global North and South. Recent catastrophic events in the USA, Spain, Libya, Germany/Belgium, and China have underscored the urgent need for enhanced early warning capabilities, particularly as climate change intensifies extreme rainfall events, even in historically low-risk regions. This thesis has addressed a critical gap in global disaster preparedness by developing and demonstrating a proof-of-concept for medium-range predictions of areas at risk of flash floods over a continuous global domain.

Three contributions: flash-flood-focused verification, data-driven modelling of imbalanced datasets, and transfer learning for global coverage

The research presented in this thesis makes three fundamental contributions to flash flood science and operational forecasting. First, through the development of a flash-flood-focused verification framework (Chapter 5), this work establishes that global numerical weather prediction models can identify areas at risk of flash floods with meaningful skill up to medium-range timescales (5 days ahead), challenging the prevailing assumption that flash flood prediction must use high-resolution forecasts to generate useful predictions. Second, the successful implementation of data-driven hydro-meteorological models (Chapter 6) demonstrates that machine learning approaches can extract predictive signals from severely imbalanced datasets, in which flash flood events are reported less than 0.3% of the time. Third, and most significantly for global applications, this research proves that models trained exclusively on high-quality observations from

data-rich regions can successfully predict flash flood risk in entirely different geographical contexts (Chapter 7).

Globally available NWP outputs and transfer learning circumvent the fundamental barrier of absent observational networks

The methodology developed here offers a pragmatic pathway toward this ambitious goal, which has particular relevance for the billions of people residing in data-scarce regions where traditional forecasting approaches remain economically or technically unfeasible. By leveraging globally available global numerical weather prediction outputs from systems like ERA5 and employing transfer learning principles, this approach circumvents the fundamental obstacle that has historically prevented global flash flood warning coverage: the absence of comprehensive observational networks in most parts of the world.

Operational protocols: high-confidence actions at day 1, preparedness activities at days 3–5

The operational implications of this research extend beyond technical achievements. Emergency managers could henceforth access flash flood risk assessments with sufficient lead time to implement graduated response protocols. Day-one forecasts, maintaining performance metrics comparable to reanalysis-based predictions, support high-confidence decisions including evacuations and resource mobilisation. Medium-range forecasts at days three to five, whilst experiencing some skill degradation, retain sufficient accuracy for preparedness activities such as pre-positioning emergency supplies, issuing public advisories, and activating monitoring protocols. This temporal stratification of forecast utility enables cost-effective risk management strategies that balance the imperative to protect lives against the economic and social costs of false alarms.

Limitations: coarse resolution, impact database biases, and absence of global verification data

Several limitations constrain the immediate operational deployment of this system and merit acknowledgement. The 31-kilometre spatial resolution inherited from ERA5 cannot fully capture the localised processes generating flash floods in small catchments, and so cannot pinpoint elevated risks that might be intrinsic to narrow valleys or urban areas. The reliance on impact databases for model training introduces systematic biases related to population density and socio-economic factors that may not transfer uniformly to all global contexts. The absence of comprehensive global verification data prevents rigorous statistical validation of model performance in most regions, necessitating continued reliance on case-study validation. These constraints, whilst significant, do not negate the fundamental advance this research represents in extending flash flood warnings to previously unprotected populations.

Four future research directions: higher-resolution models, ensemble uncertainty, global training data aggregation, and causal inference

Future research should pursue four strategic directions to enhance global flash flood prediction capabilities. First, integration with higher-resolution atmospheric models, including convection-permitting implementations, could substantially improve the representation of localised extreme rainfall events, most notably for shorter lead times. Second, ensemble approaches incorporating multiple data-driven architectures has potential to provide better uncertainty quantification for risk-based decision-making. Third, systematic aggregation of high-density, regional flash flood databases could create more globally representative training datasets. Fourth, the development of causal inference techniques could address spurious correlations identified through machine learning, ensuring predictions align with physical understanding whilst maintaining statistical accuracy.

Thesis demonstrates that universal flash flood early warning is technically feasible through NWP, machine learning, and transfer learning

The vision of universal flash flood protection, advocated by initiatives such as the UN's "Early Warnings for All" initiative, appeared technically unattainable when announced in 2022. This thesis demonstrates that the integration of global numerical weather prediction models, data-driven methodologies, and transfer learning can transform this vision into an operational reality. The successful prediction of devastating flash floods in Valencia and Zhengzhou using models trained exclusively on North American data supports the fundamental hypothesis that hydro-meteorological relationships governing flash flood generation exhibit sufficient commonality to enable knowledge transfer across continents. As climate change drives communities into unfamiliar hazard regimes, this capability to extend protective warnings beyond traditional observational boundaries becomes not merely advantageous but essential for global resilience. The proof-of-concept established through this research provides the scientific foundation for a new generation of early warning systems that can help protect vulnerable communities worldwide, ultimately contributing to the reduction of flash flood mortality and the enhancement of global disaster preparedness.

BIBLIOGRAPHY

Addor, N., H. X. Do, C. Alvarez-Garreton, G. Coxon, K. Fowler, and P. A. Mendoza, 2020: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, **65 (5)**, 712–725, https://doi.org/10.1080/02626667.2019.1683182.

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark, 2017: The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, **21 (10)**, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, URL https://hess.copernicus.org/articles/21/5293/2017/.

Adler, R. F., and Coauthors, 2018: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. *Atmosphere*, **9 (4)**, 138, https://doi.org/10.3390/atmos9040138.

Agabiirwe, C. N., P. Dambach, T. C. Methula, and R. K. Phalkey, 2022: Impact of floods on undernutrition among children under five years of age in low- and middle-income countries: a systematic review. *Environmental Health*, **21 (98)**, 1–21, https://doi.org/10.1186/s12940-022-00910-7.

Agrawal, K., and Coauthors, 2017: A Comparison of Class Imbalance Techniques for Real-World Landslide Predictions. 1–8, https://doi.org/10.1109/MLDS.2017.21.

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019: Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge*

Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2623–2631, https://doi.org/10.1145/3292500.3330701.

Al-Rawas, G., M. R. Nikoo, and M. Al-Wardy, 2024: A review on the prevention and control of flash flood hazards on a global scale: Early warning systems, vulnerability assessment, environmental, and public health burden. *International Journal of Disaster Risk Reduction*, **115**, 105 024, https://doi.org/10.1016/j.ijdrr.2024.105024.

Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger, 2013: GloFAS-global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences Discussions*, **17 (3)**, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013.

Alfieri, L., and J. Thielen, 2015: A European precipitation index for extreme rain-storm and flash flood early warning. *Meteorological Applications*, **22 (1)**, 3–13, https://doi.org/10.1002/met.1328.

Altalhan, M., A. Algarni, and M. Turki-Hadj Alouane, 2025: Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access*, **13**, 13 686–13 699, https://doi.org/10.1109/ACCESS.2025.3531662.

Alzubaidi, L., and Coauthors, 2023: A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, **10 (1)**, 46, https://doi.org/10.1186/s40537-023-00727-2.

Andrews, L., and T. E. Grantham, 2024: Strategic stream gauging network design for sustainable water management. *Nature Sustainability*, **7 (6)**, 714–723, https://doi.org/10.1038/s41893-024-01357-z.

Archer, D. R., and H. J. Fowler, 2018: Characterising flash flood response to intense rainfall and impacts using historical information and gauged data in Britain. *Journal of Flood Risk Management*, **11**, S121–S133, https://doi.org/10.1111/jfr3.12187.

Balsamo, G., A. Beljaars, K. Scipal, P. Viterbo, B. v. d. Hurk, M. Hirschi, and A. K. Betts, 2009: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System. *Journal of Hydrometeorology*, **10 (3)**, 623–643, https://doi.org/10.1175/2008JHM1068.1.

Barrett, A. I., C. Wellmann, A. Seifert, C. Hoose, B. Vogel, and M. Kunz, 2019: One step at a time: How model time step significantly affects convection-permitting simulations. *Journal of Advances in Modeling Earth Systems*, **11 (3)**, 641–658, https://doi.org/10.1029/2018MS001418.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bulletin of the American Meteorological Society*, **96 (11)**, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Bartholmes, J. C., J. Thielen, M. H. Ramos, and S. Gentilini, 2009: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, **13 (2)**, 141–153, https://doi.org/10.5194/hess-13-141-2009.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55, https://doi.org/10.1038/nature14956.

Bazo, J., R. Singh, M. Destrooper, and E. C. de Perez, 2019: Pilot experiences in using seamless forecasts for early action: The ”ready-set-go!” approach in the Red Cross. *Sub-seasonal to seasonal prediction: The gap between weather and climate forecasting*, Elsevier, 387–398, https://doi.org/10.1016/B978-0-12-811714-9.00018-8.

Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. Van Dijk, T. R. McVicar, and R. F. Adler, 2019: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bulletin of the American Meteorological Society*, **100 (3)**, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1.

Bertola, M., and Coauthors, 2023: Megafloods in Europe can be anticipated from observations in hydrologically similar catchments. *Nature Geoscience*, **16**, 982–988, https://doi.org/10.1038/s41561-023-01300-5.

Beven, J. L., A. Hagen, and R. Berg, 2022: Tropical cyclone report: Hurricane Ida. Tech. Rep. AL092021, National Hurricane Center.

Beven, K. J., 2025: A short history of philosophies of hydrological model evaluation and hypothesis testing. *WIREs Water*, **12 (1)**, e1761, https://doi.org/10.1002/wat2.1761.

Bischiniotis, K., B. van den Hurk, E. Zsoter, E. Coughlan de Perez, M. Grillakis, and J. C. Aerts, 2019: Evaluation of a global ensemble flood prediction system in Peru. *Hydrological Sciences Journal*, **64 (10)**, 1171–1189, https://doi.org/10.1080/02626667.2019.1617868.

Blöschl, G., and Coauthors, 2015: Increasing river floods: fiction or reality? *Wiley Interdisciplinary Reviews: Water*, **2 (4)**, 329–344, https://doi.org/10.1002/wat2.1079.

Borga, M., M. Stoffel, L. Marchi, F. Marra, and M. Jakob, 2014: Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows. *Journal of Hydrology*, **518 (Part B)**, 194–205, https://doi.org/10.1016/j.jhydrol.2014.05.022.

Bottazzi, M., and Coauthors, 2024: High performance computing to support land, climate, and user-oriented services: The HIGHLANDER Data Portal. *Meteorological Applications*, **31 (2)**, e2166, https://doi.org/10.1002/met.2166.

Bouallegue, Z. B., T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Monthly Weather Review*, **148 (5)**, 2049–2062, https://doi.org/10.1175/mwr-d-19-0323.1.

Bouallègue, Z. B., and D. S. Richardson, 2022: On the ROC Area of Ensemble Forecasts for Rare Events. *Weather and Forecasting*, **37 (5)**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

Braud, I., and Coauthors, 2014: Multi-scale hydrometeorological observation and modelling for flash flood understanding. *Hydrology and Earth System Sciences*, **18 (9)**, 3733–3761, https://doi.org/10.5194/hess-18-3733-2014.

Bucherie, A., C. Hultquist, S. Adamo, C. Neely, F. Ayala, J. Bazo, and A. Kruczkiewicz, 2022a: A comparison of social vulnerability indices specific to flooding in Ecuador: principal component analysis (PCA) and expert knowledge. *International Journal of Disaster Risk Reduction*, **73**, 102 897, https://doi.org/10.1016/j.ijdrr.2022.102897.

Bucherie, A., M. Werner, M. V. D. Homberg, and S. Tembo, 2022b: Flash flood warnings in context: Combining local knowledge and large-scale hydro-meteorological patterns. *Natural Hazards and Earth System Sciences*, **22 (2)**, 461–480, https://doi.org/10.5194/nhess-22-461-2022.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, **141 (693)**, 3366–3382, https://doi.org/10.1002/qj.2619.

Byaruhanga, N., D. Kibirige, S. Gokool, and G. Mkhonta, 2024: Evolution of flood prediction and forecasting models for flood early warning systems: a scoping review. *Water*, **16 (13)**, 1763, https://doi.org/10.3390/w16131763.

Casati, B., and Coauthors, 2008: Forecast verification: current status and future directions. *Meteorological Applications*, **15 (1)**, 3–18, https://doi.org/10.1002/met.52.

Castorina, G., M. T. Caccamo, V. Insinga, S. Magazù, G. Munaò, C. Ortega, A. Semprebello, and U. Rizza, 2022: Impact of the different grid resolutions of the WRF model for the forecasting of the flood event of 15 july 2020 in palermo (italy). *Atmosphere*, **13 (10)**, 1717, https://doi.org/10.3390/atmos13101717.

Cavaiola, M., F. Cassola, D. Sacchetti, F. Ferrari, and A. Mazzino, 2024: Hybrid AI-enhanced lightning flash prediction in the medium-range forecast horizon. *Nature Communications*, **15 (1188)**, 1–15, https://doi.org/10.1038/s41467-024-44697-2.

Chang, W., and X. Chen, 2018: Monthly Rainfall-Runoff Modeling at Watershed Scale: A Comparative Study of Data-Driven and Theory-Driven Approaches. *Water*, **10 (9)**, 1116, https://doi.org/10.3390/w10091116.

Chen, T., and C. Guestrin, 2016: XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, 785–794, https://doi.org/10.1145/2939672.2939785.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteorological applications*, Vol. 23, 165–181, https://doi.org/10.1002/met.1538, number: 2.

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products. *Weather and Forecasting*, **29 (2)**, 377–392, https://doi.org/10.1175/WAF-D-12-00124.1.

Clerc-Schwarzenbach, F., G. Selleri, M. Neri, E. Toth, I. van Meerveld, and J. Seibert, 2024: Large-sample hydrology – a few camels or a whole caravan? *Hydrology and Earth System Sciences*, **28 (17)**, 4219–4237, https://doi.org/10.5194/hess-28-4219-2024.

Collier, C. G., 2007: Flash flood forecasting: What are the limits of predictability? *Quarterly Journal of the Royal Meteorological Society*, **133 (622)**, 3–23, https://doi.org/10.1002/qj.29.

Corral, C., M. Berenguer, D. Sempere-Torres, L. Poletti, F. Silvestro, and N. Rebora, 2019: Comparison of two early warning systems for regional flash flood hazard forecasting. *Journal of Hydrology*, **572**, 603–619, https://doi.org/10.1016/j.jhydrol.2019.03.026.

Coughlan De Perez, E., and Coauthors, 2022: Adapting to climate change through anticipatory action: The potential use of weather-based early warnings. *Weather and Climate Extremes*, **38**, 100 508, https://doi.org/10.1016/j.wace.2022.100508.

Davis, R. S., 2001: Flash flood forecast and detection methods. *Severe Convective Storms*, 481–525, https://doi.org/10.1007/978-1-935704-06-5_12.

Dimitriadis, T., T. Gneiting, and A. I. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, **118 (8)**, e2016191 118, https://doi.org/10.1073/pnas.2016191118.

Done, J. M., G. C. Craig, S. L. Gray, and P. A. Clark, 2012: Case-to-case variability of predictability of deep convection in a mesoscale model. *Quarterly Journal of the Royal Meteorological Society*, **138 (664)**, 638–648, https://doi.org/10.1002/qj.943, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.943, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.943.

Dordevic, M., P. Mutic, and H. Kim, 2020: Flash Flood Guidance System: Response to one of the deadliest hazards. URL https://wmo.int/media/magazine-article/flash-flood-guidance-system-response-one-of-deadliest-hazards.

Doswell, C. A., 2001: Severe Convective Storms—An Overview. *Severe Convective Storms*, C. A. Doswell, Ed., American Meteorological Society, Boston, MA, 1–26.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11 (4)**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Dotzek, N., P. Groenemeijer, B. Feuerstein, and A. M. Holzer, 2009: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmospheric Research*, **93 (1-3)**, 575–586, https://doi.org/10.1016/j.atmosres.2008.10.020.

Dougherty, E., and K. L. Rasmussen, 2019: Climatology of Flood-Producing Storms and Their Associated Rainfall Characteristics in the United States. *Monthly Weather Review*, **147 (11)**, 3861–3877, https://doi.org/10.1175/MWR-D-19-0020.1.

Douglas, I., 2017: Flooding in African cities, scales of causes, teleconnections, risks, vulnerability and impacts. *International Journal of Disaster Risk Reduction*, **26**, 34–42, https://doi.org/10.1016/j.ijdrr.2017.09.024.

Douinot, A., H. Roux, P. A. Garambois, K. Larnier, D. Labat, and D. Dartus, 2016: Accounting for rainfall systematic spatial variability in flash flood forecasting. *Journal of Hydrology*, **541**, 24, https://doi.org/10.1016/J.JHYDROL.2015.08.024.

Duan, L., and Coauthors, 2022: Susceptibility assessment of flash floods: a bibliometrics analysis and review. *Remote Sensing*, **14 (21)**, 5432, https://doi.org/10.3390/rs14215432.

Duchenne-Moutien, R. A., and H. Neetoo, 2021: Climate Change and Emerging Food Safety Issues: A Review. *Journal of Food Protection*, **84 (11)**, 1884–1897, https://doi.org/10.4315/JFP-21-141.

Ebi, K. L., and Coauthors, 2021: Extreme weather and climate change: Population health and health system implications. *Annual Review of Public Health*, **42**, 293–315, https://doi.org/10.1146/annurev-publhealth-012420-105026.

ECMWF, 2016: Part IV: Physical Processes. *IFS Documentation CY41R2*, IFS Documentation, ECMWF, https://doi.org/10.21957/tr5rv27xu.

Efstratiadis, A., and D. Koutsoyiannis, 2010: One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrological Sciences Journal*, **55 (1)**, 58–78, https://doi.org/10.1080/02626660903526292.

Emerton, R. E., and Coauthors, 2016: Continental and global scale flood forecasting systems. *Wiley Interdisciplinary Reviews: Water*, **3 (3)**, https://doi.org/10.1002/wat2.1137.

Flamig, Z. L. Z. L., H. Vergara, and J. J. Gourley, 2020: The ensemble framework for flash flood forecasting (EF5) v1.2: Description and case study. *Geoscientific Model Development*, **13 (10)**, 4943–4958, https://doi.org/10.5194/gmd-13-4943-2020.

Fowler, H. J., and Coauthors, 2021: Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, **2 (2)**, 107–122, https://doi.org/10.1038/s43017-020-00128-6.

Gacu, J. G., C. E. F. Monjardin, R. G. T. Mangulabnan, and J. C. F. Mendez, 2025: Application of Artificial Intelligence in Hydrological Modeling for Streamflow Prediction in Ungauged Watersheds: A Review. *Water*, **17 (18)**, 2722, https://doi.org/10.3390/w17182722, URL https://www.mdpi.com/2073-4441/17/18/2722.

Gascón, E., L. Magnusson, T. Hewson, J. Rey, and J. Rodríguez, 2025: Extreme precipitation in Spain's Valencia region. *ECMWF Newsletter*, **183**, URL https://www.ecmwf.int/en/newsletter/183/news/extreme-precipitation-spains-valencia-region.

Gascón, E., A. Montani, and T. D. Hewson, 2024: Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. *Quarterly Journal of the Royal Meteorological Society*, **150 (762)**, 3117–3145, https://doi.org/10.1002/qj.4753.

Gaume, E., M. Borga, M. C. Llasat, S. Maouche, M. Lang, and M. Diakakis, 2016: Mediterranean extreme floods and flash floods. *The mediterranean region under climate change. A scientific update*.

Gaume, E., and Coauthors, 2009: A compilation of data on European flash floods. *Journal of Hydrology*, **367 (1-2)**, https://doi.org/10.1016/j.jhydrol.2008.12.028.

Georgakakos, K. P., 1987: Real-time flash flood prediction. *Journal of geophysical research: Atmospheres*, **92 (D8)**, 9615–9629, https://doi.org/10.1029/JD092iD08p09615.

Georgakakos, K. P., and Coauthors, 2022: The flash flood guidance system implementation worldwide: a successful multidecadal research-to-operations effort. *Bulletin of the American Meteorological Society*, **103 (3)**, E665–E679, https://doi.org/10.1175/bams-d-20-0241.1.

Gourley, J. J., and Coauthors, 2017: The FLASH project - improving the tools for flash flood monitoring and prediction across the united states. *Bulletin of the American Meteorological Society*, **98 (2)**, https://doi.org/10.1175/BAMS-D-15-00247.1.

Gouweleeuw, B. T., J. Thielen, G. Franchello, A. P. J. De Roo, and R. Buizza, 2005: Flood forecasting using medium-range probabilistic weather prediction. *Hydrology and Earth System Sciences*, **9 (4)**, 365–380, https://doi.org/10.5194/hess-9-365-2005.

Grau-Bove, J., R. Higha, S. Orr, and P. Kumar, 2024: Short note on the mapping of heritage sites impacted by the 2024 floods in Valencia, Spain. *arXiv*, https://doi.org/10.48550/arXiv.2411.08717.

Guo, Z., J. P. Leitão, N. E. Simões, and V. Moosavi, 2021: Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *Journal of Flood Risk Management*, **14 (1)**, e12 684, https://doi.org/10.1111/jfr3.12684.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian, 2014: Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences*, **18 (2)**, 463–477, https://doi.org/10.5194/hess-18-463-2014.

Gupta, S. K., and D. P. Shukla, 2023: Handling data imbalance in machine learning based landslide susceptibility mapping: a case study of Mandakini River Basin, North-Western Himalayas. *Landslides*, **20 (5)**, 933–949, https://doi.org/10.1007/s10346-022-01998-1.

Göber, M., E. Zsótér, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteorological Applications*, **15 (3)**, 359–365, https://doi.org/10.1002/met.78.

Haiden, T., and S. Duffy, 2016: Use of high-density observations in precipitation verification. *ECMWF Newsletter*, **147**, 20–25.

Haiden, T., M. Janousek, F. Vitart, Z. Ben-Bouallegue, and F. Prates, 2023: Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memoranda*, **911**.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, **87 (1)**, https://doi.org/10.1175/BAMS-87-1-33.

Hancock, J., J. M. Johnson, and T. M. Khoshgoftaar, 2022: A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data. *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, 135–142, https://doi.org/10.1109/CIC56439.2022.00028.

Hapuarachchi, H. A. P., Q. J. Wang, and T. C. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrological Processes*, **25 (18)**, https://doi.org/10.1002/hyp.8040.

Harrigan, S., and Coauthors, 2020: GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, **12 (3)**, 2043–2060, https://doi.org/10.5194/essd-12-2043-2020.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, New York, NY, https://doi.org/10.1007/978-0-387-84858-7.

He, Y., A. Bárdossy, and E. Zehe, 2011: A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, **15 (11)**, 3539–3553, https://doi.org/10.5194/hess-15-3539-2011.

Hemri, S., T. Hewson, J. Rajczak, J. Bhend, L. Moret, and M. A. Liniger, 2022: How do ecPoint precipitation forecasts compare with postprocessed multi-model ensemble predictions over Switzerland? *ECMWF Technical Memoranda*, **901**, https://doi.org/10.21957/hy89j7svk.

Henao Salgado, M. J., and J. Zambrano Nájera, 2022: Assessing Flood Early Warning Systems for Flash Floods. *Frontiers in Climate*, **4 (787042)**, 1–15, https://doi.org/10.3389/fclim.2022.787042.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Weather and Forecasting*, **31 (6)**, https://doi.org/10.1175/WAF-D-16-0093.1.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, https://doi.org/10.1002/qj.3803.

Hewson, T., 2024: Capturing extreme rainfall events. *ECMWF Newsletter*, **(178)**.

Hewson, T., F. Pillosu, E. Gascon, and M. Vučković, 2023: Post-processing ERA5 output with ecPoint. *ECMWF Newsletter*, **176**.

Hewson, T. D., and F. M. Pillosu, 2021: A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, **2 (1)**, 1–10, https://doi.org/10.1038/s43247-021-00185-9.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Weather and Forecasting*, **28 (2)**, https://doi.org/10.1175/WAF-D-12-00113.1.

Hrachowitz, M., and Coauthors, 2013: A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, **58 (6)**, 1198–1255, https://doi.org/10.1080/02626667.2013.803183.

Ibarreche, J., and Coauthors, 2020: Flash Flood Early Warning System in Colima, Mexico. *Sensors*, **20 (18)**, 5231, https://doi.org/10.3390/s20185231.

Imhoff, R. O., C. C. Brauer, K. J. van Heeringen, R. Uijlenhoet, and A. H. Weerts, 2022: Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning. *Water Resources Research*, **58 (3)**, e2021WR031 591, https://doi.org/10.1029/2021WR031591.

IPCC, 2023: Climate change 2023: Synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change [core writing team, H. Lee and J. Romero (eds.)]. *IPCC, Geneva, Switzerland*, https://doi.org/10.59327/IPCC/AR6-9789291691647.

Iqbal, J., H. Bux, and S. Sahitia, 2023: Health Consequences of Natural Disasters: An Overview of Recent Literature on Floods. *Pakistan Journal of Public Health*, **13 (4)**, 192–199, https://doi.org/10.32413/pjph.v13i4.1287.

Jasper, K., J. Gurtz, and H. Lang, 2002: Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and

forecasts with a distributed hydrological model. *Journal of Hydrology*, **267 (1-2)**, 40–52, https://doi.org/10.1016/S0022-1694(02)00138-5.

Javelle, P., D. Organde, J. Demargne, C. Saint-Martin, C. de Saint-Aubin, L. Garandeau, and B. Janet, 2016: Setting up a French national flash flood warning system for ungauged catchments based on the AIGA method. Vol. 7, 1–11, https://doi.org/10.1051/e3sconf/20160718010.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed., John Wiley & Sons.

Jonkman, S. N., A. Curran, and L. M. Bouwer, 2024: Floods have become less deadly: an analysis of global flood fatalities 1975–2022. *Natural Hazards*, **120 (7)**, 2273–2295, https://doi.org/10.1007/S11069-024-06444-0/FIGURES/8.

Juba, B., and H. S. Le, 2019: Precision-Recall versus Accuracy and the Role of Large Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33 (01)**, 4039–4048, https://doi.org/10.1609/aaai.v33i01.33014039.

Kaur, H., H. S. Pannu, and A. K. Malhi, 2019: A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.*, **52 (4)**, 79:1–79:36, https://doi.org/10.1145/3343440.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017: LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

Kiptum, A., and Coauthors, 2023: Advancing operational flood forecasting, early warning and risk management with new emerging science: Gaps, opportunities and barriers in Kenya. *Journal of Flood Risk Management*, **18 (1)**, e12 884, https://doi.org/10.1111/jfr3.12884.

Kratzert, F., M. Gauch, D. Klotz, and G. Nearing, 2024: HESS opinions: Never train a long short-term memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, **28 (17)**, https://doi.org/10.5194/hess-28-4187-2024.

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, 2018: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, **22 (11)**, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, 2019: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, **23 (12)**, https://doi.org/10.5194/hess-23-5089-2019.

Kratzert, F., and Coauthors, 2023: Caravan - A global community dataset for large-sample hydrology. *Scientific Data*, **10 (61)**, 1–11, https://doi.org/10.1038/s41597-023-01975-w.

Kumar, P., R. Bhatnagar, K. Gaur, and A. Bhatnagar, 2021: Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, **1099 (012077)**, 1–8, https://doi.org/10.1088/1757-899X/1099/1/012077.

Kumari, P., and D. Toshniwal, 2021: Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. *Journal of Cleaner Production*, **279**, 123 285, https://doi.org/10.1016/j.jclepro.2020.123285.

Lavers, D. A., S. Harrigan, and C. Prudhomme, 2021: Precipitation biases in the ECMWF integrated forecasting system. *Journal of Hydrometeorology*, **22 (5)**, 1315–1334, https://doi.org/10.1175/jhm-d-20-0308.1.

LeComte, D., 2022: U.S. Weather Highlights: 2021 Hurricane Ida, Historic Cold, Heat, Drought, and Severe Storms. *Weatherwise*, **75 (3)**, 14–23, https://doi.org/10.1080/00431672.2022.2042127.

Lee, J., D. Perera, T. Glickman, and L. Taing, 2020: Water-related disasters and their health impacts: A global review. *Progress in Disaster Science*, **8 (100123)**, 1–17, https://doi.org/10.1016/j.pdisas.2020.100123.

Leong, S. H., A. Parodi, and D. Kranzlmüller, 2017: A robust reliable energy-aware urgent computing resource allocation for flash-flood ensemble forecasting on HPC infrastructures for decision support. *Future Generation Computer Systems*, **68**, 136–149, https://doi.org/10.1016/j.future.2016.09.014.

Liang, Q., X. Xia, and J. Hou, 2016: Catchment-scale High-resolution Flash Flood Simulation Using the GPU-based Technology. *Procedia Engineering*, **154**, 975–981, https://doi.org/10.1016/j.proeng.2016.07.585.

Liao, Y., Z. Wang, X. Chen, and C. Lai, 2023: Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model. *Journal of Hydrology*, **624**, 129 945, https://doi.org/10.1016/j.jhydrol.2023.129945.

Liu, C., L. Guo, L. Ye, S. Zhang, Y. Zhao, and T. Song, 2018: A review of advances in China's flash flood early-warning system. *Natural Hazards*, **92 (2)**, 619–634, https://doi.org/10.1007/s11069-018-3173-7.

Liu, Q., M. Du, Y. Wang, J. Deng, W. Yan, C. Qin, M. Liu, and J. Liu, 2024: Global, regional and national trends and impacts of natural floods, 1990–2022. *Bulletin of the World Health Organization*, **102 (6)**, 410–420, https://doi.org/10.2471/BLT.23.290243.

Liu, Y., Y. Wang, and J. Zhang, 2012: New Machine Learning Algorithm: Random Forest. *Information Computing and Applications*, B. Liu, M. Ma, and J. Chang, Eds., Springer, Berlin, Heidelberg, 246–252, https://doi.org/10.1007/978-3-642-34062-8_32.

Luo, L., Y. Wang, Q. Li, M. Li, J. Wang, G. Zhao, and M. Ma, 2025: Exploration of the spatiotemporal characteristics and triggering factors of flash flood in China. *Ecological Indicators*, **176**, 113 698, https://doi.org/10.1016/j.ecolind.2025.113698.

Luong, T. T., J. Pöschmann, R. Kronenberg, and C. Bernhofer, 2021: Rainfall threshold for flash flood warning based on model output of soil moisture: Case study Wernersbach, Germany. *Water (Switzerland)*, **13 (8)**, 1061, https://doi.org/10.3390/w13081061.

Maqtan, R., F. Othman, W. Z. Wan Jaafar, M. Sherif, and A. El-Shafie, 2022: A scoping review of flash floods in Malaysia: current status and the way forward. *Natural Hazards*, **114 (3)**, 2387–2416, https://doi.org/10.1007/s11069-022-05486-6.

Marchi, L., M. Borga, E. Preciso, and E. Gaume, 2010: Characterisation of selected extreme flash floods in Europe and implications for flood risk management. *Journal of Hydrology*, **394 (1)**, 118–133, https://doi.org/10.1016/j.jhydrol.2010.07.017.

Marjerison, R. D., M. T. Walter, P. J. Sullivan, and S. J. Colucci, 2016: Does Population Affect the Location of Flash Flood Reports? *Journal of Applied Meteorology and Climatology*, **55 (9)**, 1953–1963, https://doi.org/10.1175/JAMC-D-15-0329.1.

Marsigli, C., and Coauthors, 2021: Review article: Observations for high-impact weather and their use in verification. *Natural Hazards and Earth System Sciences*, **21 (4)**, 1297–1312, https://doi.org/10.5194/nhess-21-1297-2021.

Mason, I., 1979: On reducing probability forecasts to yes/no forecasts. *Monthly Weather Review*, **107 (2)**, 207–211, https://doi.org/10.1175/1520-0493(1979)107<0207:ORPFTY>2.0.CO;2.

Maybee, B., and Coauthors, 2024: FOREWARNS: development and multifaceted verification of enhanced regional-scale surface water flood forecasts. *Natural Hazards and Earth System Sciences*, **24 (4)**, 1415–1436, https://doi.org/10.5194/nhess-24-1415-2024.

Mazzetti, C., D. Decremer, and C. Prudhomme, 2021: Major upgrade of the European Flood Awareness System. *ECMWF Newsletter*, **166**, https://doi.org/10.21957/32RGS58MC9.

McCabe, M. F., and Coauthors, 2017: The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, **21 (7)**, 3879–3914, https://doi.org/10.5194/hess-21-3879-2017.

Merz, B., and Coauthors, 2020: Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics*, **58 (4)**, 1–52, https://doi.org/10.1029/2020rg000704.

Merz, R., A. Miniussi, S. Basso, K.-J. Petersen, and L. Tarasova, 2022: More Complex is Not Necessarily Better in Large-Scale Hydrological Modeling: A Model Complexity Experiment across the Contiguous United States. *Bulletin of the American Meteorological Society*, **103 (8)**, E1947–E1967, https://doi.org/10.1175/BAMS-D-21-0284.1, URL https://journals.ametsoc.org/view/journals/bams/103/8/BAMS-D-21-0284.1.xml.

Meyer, J., M. Neuper, L. Mathias, E. Zehe, and L. Pfister, 2022: Atmospheric conditions favouring extreme precipitation and flash floods in temperate regions of Europe. *Hydrology and Earth System Sciences*, **26 (23)**, 6163–6183, https://doi.org/0.5194/hess-26-6163-2022.

Mitheu, F., E. Stephens, C. Petty, A. Ficchì, E. Tarnavsky, and R. Cornforth, 2023: Impact-Based Flood Early Warning for Rural Livelihoods in Uganda. *Weather, Climate, and Society*, **15 (3)**, 525–539, https://doi.org/10.1175/WCAS-D-22-0089.1.

Mitheu, F., E. Tarnavsky, A. Ficchì, E. Stephens, R. Cornforth, and C. Petty, 2025: The utility of impact data in flood forecast verification for anticipatory actions: Case studies from Uganda and Kenya. *Journal of Flood Risk Management*, **18 (1)**, e12 911, https://doi.org/10.1111/jfr3.12911.

Mosavi, A., P. Ozturk, and K. W. Chau, 2018: Flood prediction using machine learning models: Literature review. *Water*, **10 (11)**, 1536, https://doi.org/10.3390/w10111536.

Nasta, P., and Coauthors, 2025: HESS Opinions: Towards a common vision for the future of hydrological observatories. *Hydrology and Earth System Sciences*, **29 (2)**, 465–483, https://doi.org/10.5194/hess-29-465-2025.

Nearing, G., and Coauthors, 2024: Global prediction of extreme floods in ungauged watersheds. *Nature*, **627 (8004)**, 559–563, https://doi.org/10.1038/s41586-024-07145-1.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto, and H. V. Gupta, 2021: What role does hydrological science play in the age of machine learning? *Water Resources Research*, **57 (3)**, https://doi.org/10.1029/2020WR028091.

Nicótina, L., E. A. Celegon, A. Rinaldo, and M. Marani, 2008: On the impact of rainfall patterns on the hydrologic response. *Water Resources Research*, **44 (12)**, https://doi.org/10.1029/2007WR006654.

NWS, 2025: NOAA's National Weather Service - Glossary. URL https://forecast.weather.gov/glossary.php?word=flash+flood.

Oddo, P. C., J. D. Bolten, S. V. Kumar, and B. Cleary, 2024: Deep Convolutional LSTM for improved flash flood prediction. *Frontiers in Water*, **6**, 1346 104, https://doi.org/10.3389/frwa.2024.1346104.

Owens, R. G., and T. Hewson, 2018: ECMWF forecast user guide. *ECMWF*, https://doi.org/10.21957/m1cs7h.

Panigrahi, M., A. Sharma, and V. Poonia, 2025: Advancements and Challenges in Flood Modeling: A Comprehensive Review of Empirical to

Physical-Based Approaches. *Geoinformatics for Flood Risk Management*, CRC Press.

Panwar, V., and S. Sen, 2020: Disaster Damage Records of EM-DAT and DesInventar: A Systematic Comparison. *Economics of Disasters and Climate Change*, **4 (2)**, 295–317, https://doi.org/10.1007/s41885-019-00052-0.

Perez, E. C. D., and Coauthors, 2016: Action-based flood forecasting for triggering humanitarian action. *Hydrology and Earth System Sciences*, **20 (9)**, 3549–3560, https://doi.org/10.5194/hess-20-3549-2016.

Philipp, A., F. Kerl, U. Büttner, C. Metzkes, T. Singer, M. Wagner, and N. Schütze, 2016: Small-scale (flash) flood early warning in the light of operational requirements: Opportunities and limits with regard to user demands, driving data, and hydrologic modeling techniques. *IAHS-AISH Proceedings and Reports*, **373 (1)**, 201–208, https://doi.org/10.5194/piahs-373-201-2016.

Pillosu, F. M., T. Hewson, E. Stephens, C. Prudhomme, and H. Cloke, 2025a: Does a multivariate approach enhance univariate grid-to-point post-processed rainfall forecasts? A comparative analysis. *ECMWF Technical Memoranda*, **924**, 1–21.

Pillosu, F. M., T. D. Hewson, C. Prudhomme, E. Gascòn, M. Vuckovic, E. Stephens, and H. Cloke, 2025b: Bridging the scale gap: enhancing point-scale rainfall estimates through the post-processing of ERA5. *Unpublished*, 1–15.

Pillosu, F. M., and Coauthors, 2024: Can global rainfall forecasts identify areas at flash flood risk? Proof of concept for Ecuador. *ECMWF Technical Memoranda*, **917**, 1–37, https://doi.org/10.21957/8e2dd559f0, URL https://www.ecmwf.int/en/elibrary/81571-can-global-rainfall-forecasts-identify-areas-flash-flood-risk-proof-concept.

Pinos, J., and A. Quesada-Román, 2022: Flood risk-related research trends in latin america and the caribbean. *Water*, **14 (1)**, 10, https://doi.org/10.3390/w14010010.

Prakash, H., K. K. Pandey, and P. Soni, 2025: Peak discharge estimation for ungauged basins: a review. *Journal of Water and Climate Change*, **16 (11)**, 3483–3507, https://doi.org/10.2166/wcc.2025.153.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2018: CatBoost: unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 6639–6649.

Prudhomme, C., and Coauthors, 2024: Global hydrological reanalyses: The value of river discharge information for world-wide downstream applications – The example of the Global Flood Awareness System GloFAS. *Meteorological Applications*, **31 (2)**, e2192, https://doi.org/10.1002/met.2192.

Ramos Filho, G. M., V. H. Rabelo Coelho, E. da Silva Freitas, Y. Xuan, and C. das Neves Almeida, 2021: An improved rainfall-threshold approach for robust prediction and warning of flood and flash flood hazards. *Natural Hazards*, **105 (3)**, https://doi.org/10.1007/s11069-020-04405-x.

Raynaud, D., J. Thielen, P. Salamon, P. Burek, S. Anquetin, and L. Alfieri, 2015: A dynamic runoff co-efficient to improve flash flood early warning in Europe: Evaluation on the 2013 central European floods in Germany. *Meteorological Applications*, **22 (3)**, https://doi.org/10.1002/met.1469.

Robbins, J. C., and H. A. Titley, 2018: Evaluating high-impact precipitation forecasts from the Met Office Global Hazard Map (GHM) using a global impact database. *Meteorological Applications*, **25 (4)**, 548–560, https://doi.org/10.1002/met.1720.

Roberts, N., and Coauthors, 2023: IMPROVER: The New Probabilistic Postprocessing System at the Met Office. *Bulletin of the American Meteorological Society*, **104 (3)**, E680–E697, https://doi.org/10.1175/BAMS-D-21-0273.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations. *Weather and Forecasting*, **19 (5)**, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Roy, S. S., R. Chopra, K. C. Lee, C. Spampinato, and B. Mohammadi-ivatlood, 2020: Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies. *International Journal of Ad Hoc and Ubiquitous Computing*, **33 (1)**, 62–71, https://doi.org/10.1504/IJAHUC.2020.104715.

Rozemberczki, B., L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, 2022: The Shapley Value in Machine Learning. *Proceedings of the 31st International Joint Conference on Artifical Intelligence, IJCAI-ECAI 2022*, International Joint Conferences on Artificial Intelligence Organization, 5572–5579, https://doi.org/10.24963/ijcai.2022/778, URL https://www.research.ed.ac.uk/en/publications/the-shapley-value-in-machine-learning.

Saharia, M., P.-E. Kirstetter, H. Vergara, J. J. Gourley, Y. Hong, and M. Giroud, 2017: Mapping Flash Flood Severity in the United States. *Journal of Hydrometeorology*, **18 (2)**, 397–411, https://doi.org/10.1175/JHM-D-16-0082.1.

Saito, T., and M. Rehmsmeier, 2015: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, **10 (3)**, e0118 432, https://doi.org/10.1371/journal.pone.0118432.

Saleh, R. A., A. M. Al-Areeq, A. A. Al Aghbari, M. Ghaleb, M. Benaafi, N. M. Al-Areeq, and B. M. Al-Ramadan, 2024: A novel voting ensemble model empowered by metaheuristic feature selection for accurate flash flood susceptibility mapping. *Geomatics, Natural Hazards and Risk*, **15 (1)**, 2360 000, https://doi.org/10.1080/19475705.2024.2360000.

Sangati, M., and M. Borga, 2009: Influence of rainfall spatial resolution on flash flood modelling. *Natural Hazards and Earth System Sciences*, **9 (2)**, 575–584, https://doi.org/10.5194/nhess-9-575-2009.

Santos, L., and Coauthors, 2025: Machine Learning-based Hydrological Models for Flash Floods: A Systematic Literature Review. *Smart Construction and Sustainable Cities*, **3 (21)**, https://doi.org/10.1007/s44268-025-00071-9.

Sasse, L., and Coauthors, 2025: Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, **12 (1)**, 135, https://doi.org/10.1186/s40537-025-01193-8.

Schumacher, R. S., 2017: Heavy rainfall and flash flooding. *Oxford research encyclopedia of natural hazard science*, Oxford University Press, 1–42, https://doi.org/10.1093/acrefore/9780199389407.013.132.

Schwartz, C. S., 2019: Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Monthly Weather Review*, **147 (8)**, 2997–3023, https://doi.org/10.1175/MWR-D-18-0452.1.

Schwartz, C. S., and R. A. Sobash, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central and eastern United States. *Monthly Weather Review*, **147 (12)**, https://doi.org/10.1175/MWR-D-19-0115.1.

Shuvo, S. D., T. Rashid, S. K. Panda, S. Das, and D. A. Quadir, 2021: Forecasting of pre-monsoon flash flood events in the northeastern Bangladesh using coupled hydrometeorological NWP modelling system. *Meteorology and Atmospheric Physics*, https://doi.org/10.1007/s00703-021-00831-z.

Shwartz-Ziv, R., and A. Armon, 2022: Tabular data: Deep learning is not all you need. *Information Fusion*, **81**, 84–90, https://doi.org/10.1016/j.inffus.2021.11.011.

Singh, S., K. Mishra, R. Chavan, and H. L. Tiwari, 2024: Advancements and Challenges in Hydrological Modeling: A Comprehensive Review. *Hydrology and Hydrologic Modelling*, M. Pandey, N. Umamahesh, J. Das, and J. H. Pu, Eds., Springer Nature, 423–442, https://doi.org/10.1007/978-981-97-7474-6_32.

Singh, V. P., and D. A. Woolhiser, 2002: Mathematical Modeling of Watershed Hydrology. *Journal of Hydrologic Engineering*, **7 (4)**, 270–292, https://doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270).

Sofaer, H. R., J. A. Hoeting, and C. S. Jarnevich, 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, URL https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13140.

Song, T., W. Ding, J. Wu, H. Liu, H. Zhou, and J. Chu, 2020: Flash flood forecasting based on long short-term memory networks. *Water (Switzerland)*, **12 (1)**, https://doi.org/10.3390/w12010109.

Speight, L., and Coauthors, 2018: Developing surface water flood forecasting capabilities in Scotland: an operational pilot for the 2014 Commonwealth Games in Glasgow. *Journal of Flood Risk Management*, **11**, https://doi.org/10.1111/jfr3.12281.

Speight, L. J., M. D. Cranston, C. J. White, and L. Kelly, 2021: Operational and emerging capabilities for surface water flood forecasting. *Wiley Interdisciplinary Reviews: Water*, **8 (3)**, e1517, https://doi.org/10.1002/wat2.1517.

Spruce, M. D., R. Arthur, J. Robbins, and H. T. P. Williams, 2021: Social sensing of high-impact rainfall events worldwide: a benchmark comparison against manually curated impact observations. *Natural Hazards and Earth System Sciences*, **21 (8)**, 2407–2425, https://doi.org/10.5194/nhess-21-2407-2021.

Stephens, L., and J. Levi, 2024: South Sudan floods: the first example of a mass population permanently displaced by climate change? *The Conversation*, URL http://theconversation.com/south-sudan-floods-the-first-example-of-a-mass-population-permanently-displaced-by-climate-change-238461.

Sun, A. Y., and B. R. Scanlon, 2019: How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters*, **14 (7)**, 073 001, https://doi.org/10.1088/1748-9326/ab1b7d.

Tapiador, F. J., R. Roca, A. D. Genio, B. Dewitte, W. Petersen, and F. Zhang, 2019: Is Precipitation a Good Metric for Model Performance? *Bulletin of the American Meteorological Society*, **100 (2)**, 223–233, https://doi.org/10.1175/BAMS-D-17-0218.1.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, **106 (D7)**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo, 2009: The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, **13 (2)**, 125–140, https://doi.org/10.5194/hess-13-125-2009.

Todini, E., 2011: History and perspectives of hydrological catchment modelling. *Hydrology Research*, **42 (2-3)**, 73–85, https://doi.org/10.2166/nh.2011.096.

Tripathy, S. S., H. Vittal, S. Karmakar, and S. Ghosh, 2020: Flood risk forecasting at weather to medium range incorporating weather model, topography, socio-economic information and land use exposure.

Advances in Water Resources, **146**, 103 785, https://doi.org/10.1016/j.advwatres.2020.103785.

Tsonevsky, I., C. A. Doswell, and H. E. Brooks, 2018: Early warnings of severe convection using the ECMWF extreme forecast index. *Weather and Forecasting*, **33 (3)**, https://doi.org/10.1175/WAF-D-18-0030.1.

UN, 2022: Early Warnings for All. URL https://www.un.org/en/climatechange/early-warnings-for-all.

Vannitsem, S., and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Villaça, C., P. P. Santos, and J. L. Zêzere, 2025: Modeling rainfall input variability for flash floods in Portugal: the influence of predisposing factors. *Geomatics, Natural Hazards and Risk*, **16 (1)**, 2462 179, https://doi.org/10.1080/19475705.2025.2462179.

Wang, N., L. Lombardo, M. Tonini, W. Cheng, L. Guo, and J. Xiong, 2021: Spatiotemporal clustering of flash floods in a changing climate (China, 1950-2015). *Natural Hazards and Earth System Sciences*, **21 (7)**, https://doi.org/10.5194/nhess-21-2109-2021.

Wang, Y., and H. A. Karimi, 2022: Impact of spatial distribution information of rainfall in runoff simulation using deep learning method. *Hydrology and Earth System Sciences*, **26 (9)**, https://doi.org/10.5194/HESS-26-2387-2022.

Wen, Y., T. Schuur, H. Vergara, and C. Kuster, 2021: Effect of precipitation sampling error on flash flood monitoring and prediction: Anticipating operational rapid-update polarimetric weather radars. *Journal of Hydrometeorology*, **22 (7)**, https://doi.org/10.1175/JHM-D-19-0286.1.

Wikipedia, 2025: 2024 Afghanistan–Pakistan floods. URL https://en.wikipedia.org/w/index.php?title=2024_Afghanistan%E2%80%93Pakistan_floods&oldid=1266577064.

Wilks, D. S., 2020: *Statistical Methods in the Atmospheric Sciences*. 4th ed., Elsevier, https://doi.org/10.1016/C2017-0-03921-6.

WMO, 2017: WMO Guidelines on the Calculation of Climate Normals. Tech. Rep. 1203, World Meteorological Organization, Geneva, 1–29 pp.

WMO, W. M. O., 2025: State of the climate 2024. URL https://wmo.int/publication-series/state-of-global-climate-2024.

Wohl, E., and K. B. Lininger, 2022: Hydrology and discharge. *Large rivers: Geomorphology and management, second edition*, Wiley Online Library, 42–75, https://doi.org/10.1002/9781119412632.ch3.

Wyatt, F., J. Robbins, and R. Beckett, 2023: Investigating bias in impact observation sources and implications for impact-based forecast evaluation. *International journal of disaster risk reduction*, **90**, 103 639, https://doi.org/10.1016/j.ijdrr.2023.103639.

Wyatt, F., J. Robbins, and S. Eaton, 2024: Implementing a routine and standard approach for the automatic collection of socio-economic impact observations for impact-based forecasting and warning. *International Journal of Disaster Risk Reduction*, **110**, 104 608, https://doi.org/10.1016/j.ijdrr.2024.104608.

Wyngaard, J. C., 2004: Toward Numerical Modeling in the “Terra Incognita”. *Journal of the Atmospheric Sciences*, **61 (14)**, 1816–1826, https://doi.org/10.1175/1520-0469(2004)061<1816:TNMITT>2.0.CO;2.

Xu, S., Y. Song, and X. Hao, 2022: A Comparative Study of Shallow Machine Learning Models and Deep Learning Models for Landslide Susceptibility Assessment Based on Imbalanced Data. *Forests*, **13 (11)**, 1908, https://doi.org/10.3390/f13111908.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart, 2008: Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resources Research*, **44 (5)**, https://doi.org/10.1029/2007WR005940.

Yin, J., and Coauthors, 2023: Flash floods: why are more of them devastating the world’s driest regions? Setting the agenda in research. *212* | *Nature* |, **615**, URL https://gfp.jrc.ec.europa.eu.

Ying, X., 2019: An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, **1168 (2)**, 022 022.

Zanchetta, A. D. L., and P. Coulibaly, 2020: Recent advances in real-time pluvial flash flood forecasting. *Water (Switzerland)*, **12 (2)**, https://doi.org/10.3390/w12020570.

Zanchetta, A. D. L., P. Coulibaly, and V. Fortin, 2022: Forecasting high-flow discharges in a flashy catchment using multiple precipitation

estimates as predictors in machine learning models. *Hydrology*, **9 (12)**, 216, https://doi.org/10.3390/hydrology9120216.

Zeman, C., N. P. Wedi, P. D. Dueben, N. Ban, and C. Schär, 2021: Model intercomparison of COSMO 5.0 and IFS 45r1 at kilometer-scale grid spacing. *Geoscientific Model Development [preprint]*, **2021**, https://doi.org/10.5194/gmd-2021-31.

Zhai, X., L. Guo, R. Liu, and Y. Zhang, 2018: Rainfall threshold determination for flash flood warning in mountainous catchments with consideration of antecedent soil moisture and rainfall pattern. *Natural Hazards*, **94 (2)**, https://doi.org/10.1007/s11069-018-3404-y.

Zhang, S., Y. Wang, and G. Wu, 2022: Earthquake-Induced Landslide Susceptibility Assessment Using a Novel Model Based on Gradient Boosting Machine Learning and Class Balancing Methods. *Remote Sensing*, **14 (23)**, 5945, https://doi.org/10.3390/rs14235945.

Zhang, Y., and Coauthors, 2024: Impact of floods on the environment: A review of indicators, influencing factors, and evaluation methods. *Science of The Total Environment*, **951**, 175 683, https://doi.org/10.1016/j.scitotenv.2024.175683.

Zhang, Z., Q. Zhang, V. P. Singh, and P. Shi, 2018: River flow modelling: comparison of performance and evaluation of uncertainty using data-driven models and conceptual hydrological model. *Stochastic Environmental Research and Risk Assessment*, **32 (9)**, 2667–2682, https://doi.org/10.1007/s00477-018-1536-y.

Zhao, G., R. Liu, M. Yang, T. Tu, M. Ma, Y. Hong, and X. Wang, 2022: Large-scale flash flood warning in China using deep learning. *Journal of Hydrology*, **604**, https://doi.org/10.1016/j.jhydrol.2021.127222.

Zhao, Y., X. Wu, W. Zhang, P. Lan, G. Qin, X. Li, and H. Li, 2025: A deep learning-based probabilistic approach to flash flood warnings in mountainous catchments. *Journal of Hydrology*, **652**, 132 677, https://doi.org/10.1016/j.jhydrol.2025.132677.

Zhou, D.-W., Z.-W. Cai, H.-J. Ye, D.-C. Zhan, and Z. Liu, 2025: Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. *International Journal of Computer Vision*, **133 (3)**, 1012–1032, https://doi.org/10.1007/s11263-024-02218-0.

Zsoter, E., H. Cloke, E. Stephens, P. d. Rosnay, J. Muñoz-Sabater, C. Prudhomme, and F. Pappenberger, 2019: How Well Do Operational Numerical Weather Prediction Configurations Represent Hydrology? *Journal of Hydrometeorology*, **20 (8)**, 1533–1552, https://doi.org/10.1175/JHM-D-18-0086.1.

Šakić Trogrlić, R., M. van den Homberg, M. Budimir, C. McQuistan, A. Sneddon, and B. Golding, 2022: Early warning systems and their role in disaster risk reduction. *Towards the "perfect" weather warning: bridging disciplinary gaps through partnership and communication*, Springer International Publishing, 11–46, https://doi.org/https://doi.org/10.1007/978-3-030-98989-7_2.