Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting East Africa's rainfall extremes with calibrated AI post-processing systems

Shruti Nath

shruti.nath@physics.ox.ac.uk

University of Oxford

David Koros

Kenya Meteorological Service Authority

Fenwick C. Cooper

University of Oxford

David MacLeod

Cardiff University

Hannah Kimani

Kenya Meteorological Service Authority

Zacharia Mwai

Kenya Meteorological Service Authority

Christine Maswi

Kenya Meteorological Service Authority

Bernard Chanzu

Kenya Meteorological Service Authority

Asaminew Teshome

Ethiopian Meteorological Institute

Bekalu Tamene

Ethiopian Meteorological Institute

Bekele Kebebe

Ethiopian Meteorological Institute

Samrawit Abebe

Ethiopian Meteorological Institute

Masilin Gudoshava

IGAD Climate Prediction and Applications Center

Ahmed Amdihun

IGAD Climate Prediction and Applications Center

Isaac Obai

World Food Programme

Maurine Ambani

World Food Programme
Mark Arango
World Food Programme
Jesse Mason
World Food Programme
Matthew Chantry
European Centre for Medium-Range Weather Forecasts
Florian Pappenberger
European Centre for Medium-Range Weather Forecasts
Antje Weisheimer
University of Oxford
Tim Palmer
University of Oxford

Article

Keywords: Rainfall, ML/AI, extremes, Relative Economic Value, Early Warning Systems

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10487940/v1

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

Additional Declarations: No competing interests reported.

Predicting East Africa's rainfall extremes with calibrated AI post-processing systems

Shruti Nath[1*], David Koros[2†], Fenwick C. Cooper[1†], David MacLeod[3], Hannah Kimani[2], Zacharia Mwai[2], Christine Maswi[2], Bernard Chanzu[2], Asaminew Teshome[4], Bekalu Tamene[4], Bekele Kebebe[4], Samrawit Abebe[4], Masilin Gudoshava[5], Ahmed Amdihun[5], Isaac Obai[6], Maurine Ambani[6], Mark Arango[6], Jesse Mason[6], Matthew Chantry[7], Florian Pappenberger[7], Antje Weisheimer[1,7,8], Tim Palmer[1]

[1*]Department of Physics, University of Oxford, Oxford, U.K..
[2]Kenya Meteorological Department, Nairobi, Kenya.
[3]School of Earth and Environmental Sciences, University of Cardiff, Cardiff, U.K..
[4]Ethiopian Meteorological Institute, Addis Ababa, Ethiopia.
[5]Climate Prediction and Application Centre, Inter-Governmental Authority on Development (ICPAC), Nairobi, Kenya.
[6]UN WFP, Rome, Italy.
[7]European Centre for Medium-range Weather Forecasting, Reading, U.K..
[8]National Centre for Atmospheric Sciences, Leeds, U.K..

*Corresponding author(s). E-mail(s): shruti.nath@physics.ox.ac.uk;
[†]These authors contributed equally to this work.

Abstract

Accurate and reliable rainfall forecasts are crucial for Early Warning Systems (EWS). Artificial Intelligence approaches, which post-process global physical forecasts, enable low-cost ensemble generation, extending skill to medium-range lead times and offering scalable systems to inform EWS particularly under resource-constrained settings. However, assessing their practical value in representing user-relevant rainfall extremes and guiding decisions for EWS remains

critical. We present a low-cost, uncertainty quantification method that calibrates ensemble forecasts to probabilities of exceedances for rainfall extremes. Applied to raw and post-processed rainfall forecasts over East Africa, sub-regional analysis shows the threshold-calibrated, post-processed forecasts better reflecting the observed chances of extremes at higher rainfall thresholds. Case-study results show encouraging added value from post-processed forecasts at longer lead times. Forecast skill decomposed across rainfall and probability thresholds highlights the ranges of rainfall and risk tolerances where most actionable improvement arises, whether using satellite or rain-gauge data as reference. This tailored evaluation helps guide forecast adoption for EWS.

Keywords: Rainfall, ML/AI, extremes, Relative Economic Value, Early Warning Systems

1 Introduction

Operational weather centres worldwide are undergoing rapid transformation driven by advances in machine learning–based models for global weather forecasting [1–4]. Particularly for resource-constrained settings, machine-learning based models hold a competitive edge over traditional physical models by requiring modest computational infrastructure to run on, enabling low-cost probabilistic forecast generation for rapid and timely assessment of extreme weather risks. For vulnerable regions such as East Africa, where livelihoods and economies are affected by rainfall extremes and their related hazards [5, 6], these advances promise scalable and accessible forecasting systems that retain accuracy within the critical time windows needed for Early Warning Systems and Anticipatory Action. This comes at an important moment following the formal introduction of the Early Warnings for All Initiative at the COP27 in 2022 [7], aiming for the protection of lives and livelihoods from extreme climate hazards by 2027. However, current global machine-learning models operate at around 25 km resolution limiting their utility towards informing local-level decisions. Particularly for rainfall prediction over East Africa, where the main drivers are deep convective systems [8, 9], resolution has been shown to affect the quality of short- (12- to 72- hours) and medium- (three- to ten- days) range forecasts [10]. Moreover, most machine-learning based models are trained on reanalysis datasets [1, 4] or coarsened global and regional rainfall satellite products [11–13], making it difficult to immediately assess their value in representing local rainfall extremes over the region and ultimately towards Early Warning Systems and Anticipatory Action.

Artificial Intelligence (AI) based generative methods have shown promise in correcting physical model output to yield more accurate, high-resolution forecasts [14–16]. Such approaches offer easy uptake within resource-constrained settings by building upon existing infrastructure and allowing targeted training on the accepted "ground truth" datasets. They thus enable preliminary assessment of the practical utility of machine-learning based forecasts by providing specialised, regional forecasts of impact-relevant variables needed for decision-making. Here, we consider an AI-based

rainfall post-processing system prototyped within Kenyan and Ethiopian meteorological centres alongside the Greater Horn of Africa's Climate Prediction and Application Centre (ICPAC) since 2023. The system consists of a conditional Generative Adversarial Network (cGAN) that post-processes global, operational ensemble forecasts from ECMWF's Integrated Forecast System (IFS) [17] to produce more accurate, regional 24-h accumulated, probabilistic rainfall forecasts at one to five days in advance and at a 10 km resolution [18, 19]. To date, it remains the only AI-based post-processing system routinely used by regional and local meteorological centres in East Africa to produce corrected, probabilistic rainfall forecasts. This presents a unique testbed to demonstrate the practical utility of AI post-processing systems towards Early Warning Systems and Anticipatory Action.

Evaluating the utility of forecasts towards Early Warning Systems and Anticipatory Action requires focussing on their added value towards decision-making. While typical model evaluations focus on global skill metrics such as the Continuous Ranked Probability Score (CRPS) [20], such metrics do not automatically translate to how useful forecasts are in triggering timely and effective actions. Here, we extend traditional forecast evaluation towards decision-centric metrics that first consider user-defined extreme event thresholds and then account for associated risk tolerances, which collectively inform the decision-making process of whether or not to act. A first step towards this requires translation of the forecasted rainfall distributions to single probabilities of exceedances for user-defined extreme thresholds. Simply doing this by calculating empirical ensemble probabilities would be insufficient as each ensemble member is a discrete sample from the forecasted distribution, making the representation of extremes highly sensitive to ensemble size due to their occurrence in the sparser tail regions of the distribution [21]. We therefore propose an uncertainty quantification step that further calibrates hoth raw and post-processed ensemble forecasts from focussing on the full rainfall distribution, to specialising on individual user-defined extreme rainfall thresholds.

This study introduces an Ensemble Logistic Regression (ELR) uncertainty quantification step (Figure 1) that calibrates ensemble rainfall forecasts directly to the probability that rainfall exceeds a user-defined threshold. ELR applies a spline transform to all ensemble members to approximate the underlying forecast distribution, followed by logistic regression; this yields an interpretable alternative to traditional AI fine-tuning approaches and calibrates to probabilities without compressing the ensemble to a mean/variance summary. By combining this threshold-specific calibration with decision-oriented value metrics, our framework identifies when and where AI-enhanced forecasts are most useful for Early Warning Systems, and we make the evaluation results available alongside the forecasts to guide warning issuance under varying risk contexts.

The remainder of this paper is structured as follows. Section 2 evaluates the regional performance of ELR-based calibration on both raw and post-processed forecasts across user-relevant rainfall thresholds, considering forecast skill (Section 2.1), reliability (Section 2.2), and resolution (Section 2.3). Section 3 presents more detailed analysis of the forecast value over two case-study regions (Nairobi–Kiambu–East Kajiado, Kenya; Konso–Moyale, Ethiopia) and rain-gauges selected in collaboration with

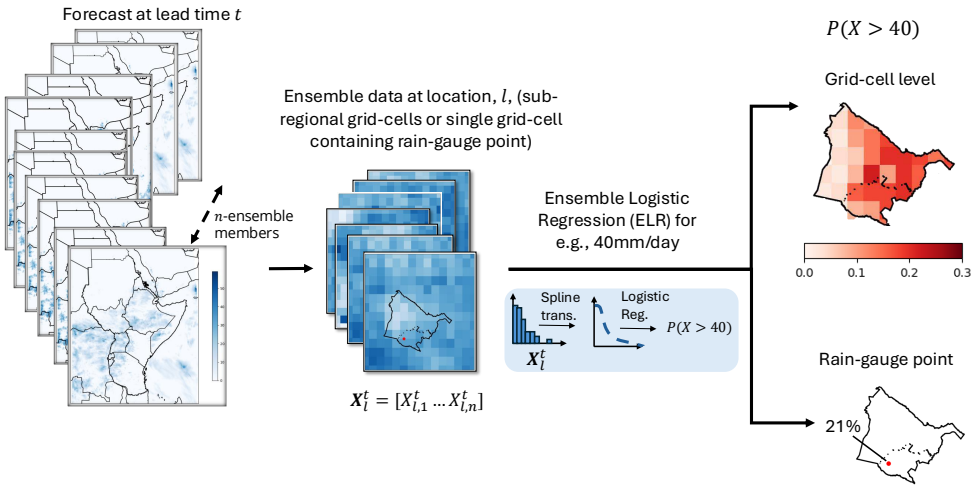


Fig. 1 Framework for calibration of ensemble output with Ensemble Logistic Regression (ELR). ELR is trained on specific sub-regions within the forecast domain. It takes ensemble, grid-cell level data as input and can be trained to predict probabilities of exceedance for a given threshold (40 mm/day shown above) at a grid-cell level or, if rain-gauge data is available for training, for that rain-gauge point.

local meteorologists. Section 4 provides conclusions and further discussion. Methodological details of the ELR and the evaluation framework can be found in Section 5.

2 Analysis over regional climate zones

We demonstrate the added value of Ensemble Logistic Regression (ELR)-based calibration to directly obtain individual probabilities of threshold exceedance. The ELR is applied to both raw (IFS+ELR) and post-processed (IFS+cGAN+ELR) forecasts, enabling a direct comparison of the calibration benefit on each. To assess the contributions of calibration and post-processing, we additionally evaluate the uncalibrated IFS and IFS+cGAN outputs, as well as the IFS calibrated with the well-established Isotonic Distributional Regression [22, 23] (IFS+IDR) as a further reference.

Separate ELR and IDR models are fitted to grid-cell level data for individual climate zones within the Greater Horn of Africa. Climate zones are chosen in a similar fashion to Köppen-Geiger climate zones [24] by considering elevation, lakes and total annual rainfall received from 2018-2021, with the additional criteria that each zone contains a sufficient number of samples such that drier regions encompass larger areas to compensate for inadequate rainfall (see Figure B2). Rainfall thresholds of 20-60 mm/day at 10 mm/day steps are fitted for. These represent the medium (20 mm/day) to severe (60 mm/day) as specified by local meteorologists and WMO standards. When training the ELR and IDR, we use the same train (2018-2020) and test (2021,2023 and

2024) year subsets as those used for the cGAN, ensuring that all evaluation results shown are largely out-of-sample across all methods. Details of the IFS, cGAN and truth rainfall datasets used can be found in Section 5.1.

We compare the models by first assessing their reductions in cross-entropy loss relative to the raw IFS ensemble (Section 2.1), then examining their reliability diagrams to assess forecast calibration (Section 2.2), and finally categorising their reliability and resolution to summarise their potential operational utility following [25] (Section 2.3).

2.1 Reduction in cross-entropy loss relative to raw IFS

Figure 2 provides the change in cross-entropy loss (negative log-likelihood of a Bernoulli distribution and a typical metric used to investigate classification error) relative to empirical IFS as a function of rainfall threshold (a lower value indicates greater skill improvement). For ease of visualisation the calculated grid-cell level cross-entropy losses are aggregated over four distinct climate zones that constitute the Greater Horn of Africa region: the Southern Indian Ocean; the Tropical Savannah, encompassing the Congo Basin and southern Tanzania; the Highlands, comprising the Ethiopian Highlands, Mt. Kenya, Mt. Kilimanjaro, Mt. Elgon, and the Rwenzori mountain ranges; and Warm semi-arid and oceanic regions, covering lower-elevation areas that still receive significant rainfall, along with parts of the Northern Indian Ocean. These zones are shaded in teal within the maps displayed at the bottom of Figure 2.

Across one, three and five day lead times, all calibration methods display a notable improvement over IFS. Apart from the Southern Indian Ocean which shows increasing improvement at thresholds greater than 40 mm/day, largest improvements are achieved at 20-30 mm/day across all forecasting systems and lead times. Across most regions, the IDR-based calibration (IFS+IDR) shows competitive performance to IFS+cGAN and apart from the Highlands, does better than IFS+cGAN at thresholds greater than 40 mm/day. This suggests that the cGAN post-processing primarily improves the body of the rainfall distribution, whereas IDR, despite calibrating the full distribution simultaneously, retains greater skill in the tails.

Within bounds of uncertainty, the ELR-based IFS+ELR and IFS+cGAN+ELR provide competitive performance to IFS+IDR, with both significantly outperforming the other methods over the Southern Indian Ocean. Below 30 mm/day, however, IFS+ELR performs worse than the other methods, which show broadly comparable skill. This suggests that the value of ELR calibration depends on the baseline skill of the input forecast: since IFS+cGAN already displays substantial skill improvement across thresholds, applying ELR on top yields a larger overall benefit than applying ELR to the raw IFS alone. Nonetheless, the added benefit from ELR is modest at lower rainfall thresholds, where IFS+cGAN already performs well, but becomes more pronounced at higher thresholds — indicating that ELR is particularly effective as an efficient, targeted calibration step for the extreme thresholds that are sparser represented in training data.

Over the warm semi-arid, oceanic, and highland areas most relevant within the political boundaries of the Greater Horn of Africa, the ELR-based approaches outperform IFS+IDR well outside the range of the uncertainty bars at rainfall thresholds ¿40 mm/day at one-day lead times. This advantage persists at three- and five-day

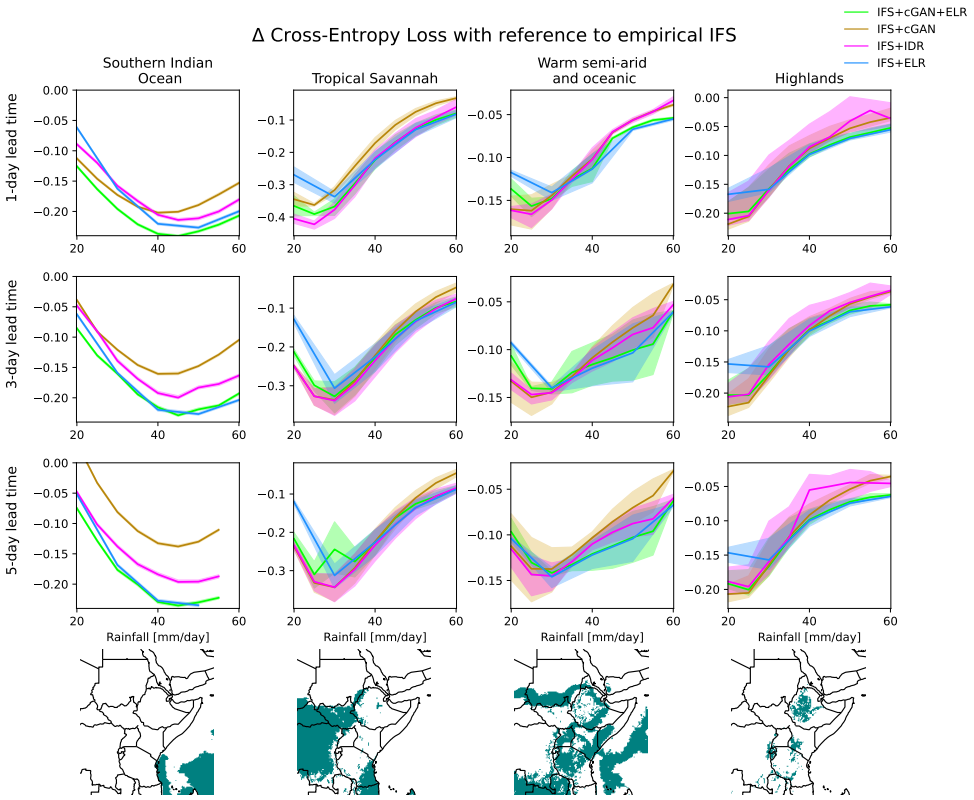


Fig. 2 Changes in cross-entropy loss relative to those obtained from empirical IFS as a function of rainfall threshold for climate zones grouped according to Southern Indian Ocean, Tropical Savannah, Highlands and Warm semi-arid and oceanic, (see bottom panel where the exact regions are shaded in teal). Changes in cross-entropy loss are provided for IFS calibrated with both Ensemble Logistic Regression (IFS+ELR) and Isotonic Distributional Regression (IFS+IDR), alongside the AI-based IFS+cGAN and ELR-calibrated IFS+cGAN+ELR approaches. Uncertainty ranges indicate the 95% confidence interval as obtained from bootstrap resampling done 1000 times.

lead times for the highland areas, but not for the warm semi-arid and oceanic areas, where performance is not significantly better than IFS+IDR. Nonetheless, these results underscore the value of threshold-specific calibration for heavy rainfall events, particularly in areas of the Greater Horn of Africa most relevant to Early Warning Systems.

2.2 Regional reliability diagrams

To assess whether the observed skill gain translates to improved forecast reliability, we inspect the reliability diagrams constructed over grid-cells comprising each climate zone. We employ the Consistency, Optimality and Reproducibility (CORP) [22] approach in constructing the reliability diagrams, which non-parametrically groups

predicted probabilities into bins within which the corresponding observed frequency of events is calculated (see Section 5.5 for further details). This allows assessment of how well predicted probabilities align with observed event frequencies, where perfect reliability corresponds to a one-to-one match. Figure 3 provides the reliability diagrams obtained over key climate zones. Climate zones are different from those shown in Figure 2 as they were selected on the basis of their relevance to decision-making within the political boundaries of the Greater Horn of Africa (see bottom panel where exact regions are shaded in teal within the maps provided). We focus on IFS+ELR and IFS+cGAN+ELR as these consistently have the best skill for the regions under consideration at thresholds above 30 mm/day. Note that the displayed rainfall thresholds vary for each region (see column label) and that the first column corresponds to the arid areas of northern Kenya, Somalia, and North Sudan, which receive very little rainfall such that the displayed threshold of 15 mm/day is already considered "heavy" rainfall.

Across all regions and lead times, at least one forecasting system retains reliability away from a flat climatological frequency, demonstrating resolution beyond a simple climatological prediction. For the arid to semi-arid regions (first and third columns) at one- and three-day lead times, IFS+cGAN+ELR exhibits a reliability curve split between low and high probabilities, indicating a sharp discrimination between non-occurrence and high-probability occurrence. The same pattern holds for IFS+ELR, except in the semi-arid regions at one-day lead time. At five-day lead time, however, IFS+cGAN+ELR improves to discriminate across a wider range of occurrence probabilities, whereas IFS+ELR degrades to showing little to no resolution.

Over the Highlands (second column), IFS+ELR displays better reliability at one-day lead times, while IFS+cGAN+ELR performs better at three- and five-day lead times, although both systems assign higher probabilities of occurrence to low-frequency events. Over the Tropical Savannah (fourth column) at one-day lead times, both IFS+cGAN+ELR and IFS+ELR show a similar tendency to assign higher probabilities of occurrence to low-frequency events. At three- to five-day lead times, IFS+cGAN+ELR shifts from a split between non-occurrence and high-probability occurrence towards more under-confident forecasts, in which even high-frequency events are assigned lower probabilities. IFS+ELR, by contrast, shows predicted probabilities concentrated within 0.1–0.5 at both lead times, albeit with greater uncertainty at five-day lead time.

2.3 Categories of forecast reliabiltiy

While skill scores and their decompositions are valuable for quantitatively assessing the added value of a forecasting system, they remain somewhat abstract for everyday practitioners and end-users who simply want to know whether a forecast is useful. In practical terms, the utility of a forecast reflects its ability to distinguish the likelihood of an event beyond what a simple climatological prediction can offer. To more systematically assess this across all regions, lead times and thresholds, we turn to the reliability categorisation framework of [25], described further in Section 5.5. Categories of reliability are constructed using empirical reliability and resolution terms obtained from the reliability analysis conducted across all climate zones, lead times and rainfall

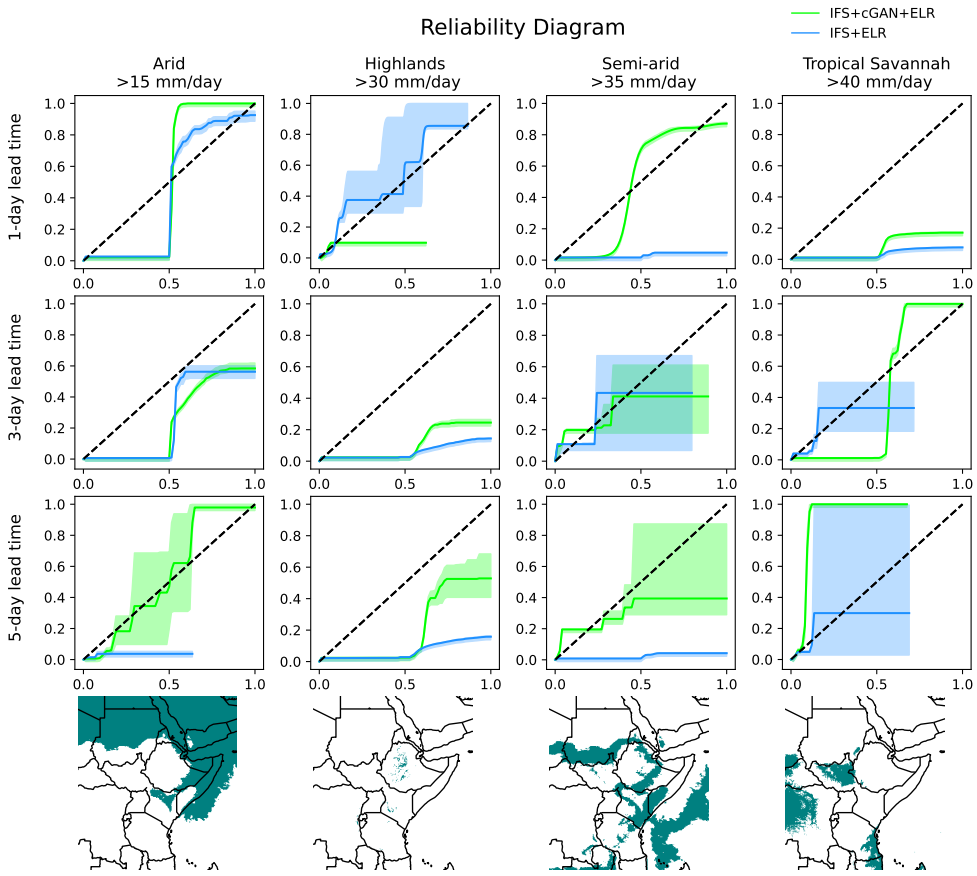


Fig. 3 Reliability diagrams for key climate zones within the Greater Horn of Africa (see bottom panel where the exact regions are shaded in teal) at >15 mm/day (first column), >20 mm/day (second column), >25 mm/day (third column) and >40 mm/day (fourth column). Results are shown for IFS+ELR and IFS+cGAN+ELR. Bins are chosen by leveraging monotonicity constraints of the reliability diagram and using a kernel-based isotonic regression.

thresholds. The reliability term is defined as the mean distance from perfect reliability over all bins (lower is better), while the resolution term is the mean distance from the climatological event frequency over all bins (higher is better). A forecast is categorised as "useful" if it has more resolution than reliability error. If the reliability error exceeds the resolution, the forecast is categorised as "marginally useful" provided it retains some resolution (i.e., resolution greater than zero), and as "not useful" otherwise. Figure 4 provides the obtained categories of reliability at one, three and five day lead times. For each region and lead time, the highest rainfall threshold for which at least one of the forecasting systems has a "useful" forecast is shown (unless this does not exist, in which case "not useful" to "marginally useful" is shown instead).

Panel (a) shows the geographical distribution of reliability, highlighting a strong improvement in forecast utility for the arid Northern Kenya, Somalia and North Sudan regions from IFS+cGAN+ELR at five day lead times, as well as for the warm semi-arid areas in southern Kenya, Ethiopia and southern North Sudan at both one and five day lead times, consistent with Figure 3. Conversely, some areas in the Ethiopian highlands show "useful" IFS+ELR at one-day lead times whereas those of IFS+cGAN+ELR are "marginally useful", again consistent with Figure 3. Most other regions remain "marginally useful" across both systems and lead times. Panel (b) of Figure 4 summarizes the number of sub-regions falling within each category of reliability at one, three and five day lead times, and shows some improvement from "not useful" to "useful" within the IFS+cGAN+ELR as opposed to IFS+ELR systems at one and five day lead times, whereas three-day lead times show similar performance between the two systems.

Overall, these results indicate that the combination of IFS+cGAN systems and ELR post-processing yields competitive performance, across a range of rainfall thresholds and specifically at longer lead times, for the areas within the Greater Horn of Africa most relevant towards EWS. Nevertheless, localised exceptions — such as the Ethiopian Highlands at one-day lead times — suggest that no single combination consistently outperforms across all regions, rainfall thresholds and lead times.

3 Case study and rain-gauge analysis of the added forecast value

The reliability categorisation establishes the rainfall thresholds at which forecasts can be considered useful, but within Early Warning Systems and Anticipatory Action, decision-makers must additionally translate what predicted probability levels for a given rainfall threshold warrant triggering action. To evaluate the added value of ELR-based calibration in this context, we consider Relative Economic Value (REV) frameworks as described in Section 5.5. To start with, we assess the value of ELR-calibrated forecasts in predicting grid-cell level exceedance probabilities for 24-hour accumulated rainfall at two case-study regions selected in collaboration with local meteorological departments (see Figure A1 for details).

The first region encompasses Nairobi, Kiambu, and East Kajiado in central Kenya, a mountainous area that typically receives rainfall exceeding 30 mm/day during the rainy seasons. The second covers Konso and Moyale in southern Ethiopia, situated within the drier lowlands where flash flooding can be triggered by rainfall as low as 20 mm/day. It should be noted that Konso and Moyale are two disjoint regions (refer to Figure A1), however were chosen due to relevance for warning dissemination within the Ethiopian Meteorological Institute. Both regions experienced significant flooding and flash flooding during the test period of 2021, 2023 and 2024, making them well-suited for evaluating forecast utility under operationally relevant conditions. Reflecting the distinct climatological characteristics of each region, we examine rainfall thresholds of 40–60 mm/day at 5 mm/day intervals for the Kenyan case study, and 25–40 mm/day at 5 mm/day intervals for the Ethiopian case study. We again focus on the IFS+ELR and

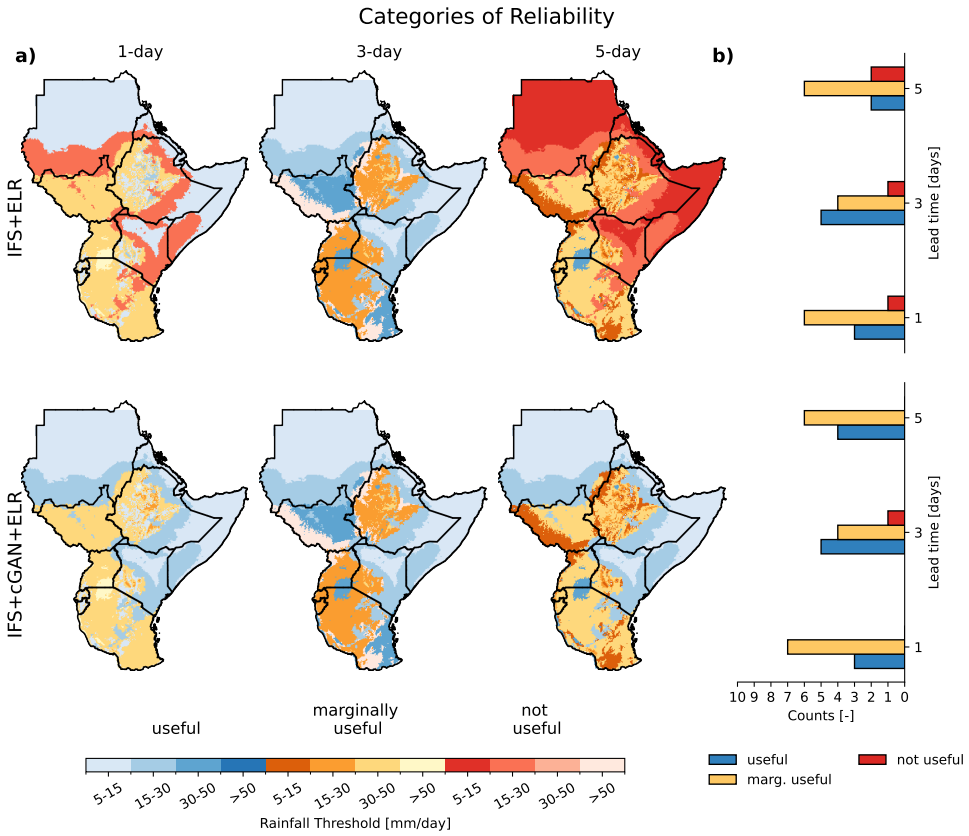


Fig. 4 Categories of reliability from ELR-based calibration of IFS and IFS+cGAN at one, three and five day lead times. For each region and lead time, the highest rainfall threshold for which at least one of the forecasting systems has a useful forecast is shown (unless this does not exist, in which case not useful to marginally useful is shown instead). Panel (a) provides maps of categories of reliability according to rainfall zones (see Figure B2). Panel (b) summarizes the maps into counts of the number of rainfall zones within each category of reliability.

IFS+cGAN+ELR forecasting systems, and the relative improvement in cross-entropy loss obtained by both are provided in Figure C3.

Figure 5 presents REV curves across different rainfall thresholds for both case-study regions at one- and five-day lead times, representative of the shortest and longest lead times available for both IFS+ELR and IFS+cGAN+ELR systems. At each probability of threshold exceedance (interchangeable with cost-loss ratio or risk level), the REV places the forecast on a scale ranging from zero — corresponding to a default forecast that always or never predicts the event — to one, representing an ideal forecast that perfectly captures all events with no false alarms or misses. Significance of the Area under the REV curves are also calculated and reported as p-values (see Section 5.5 for further details.

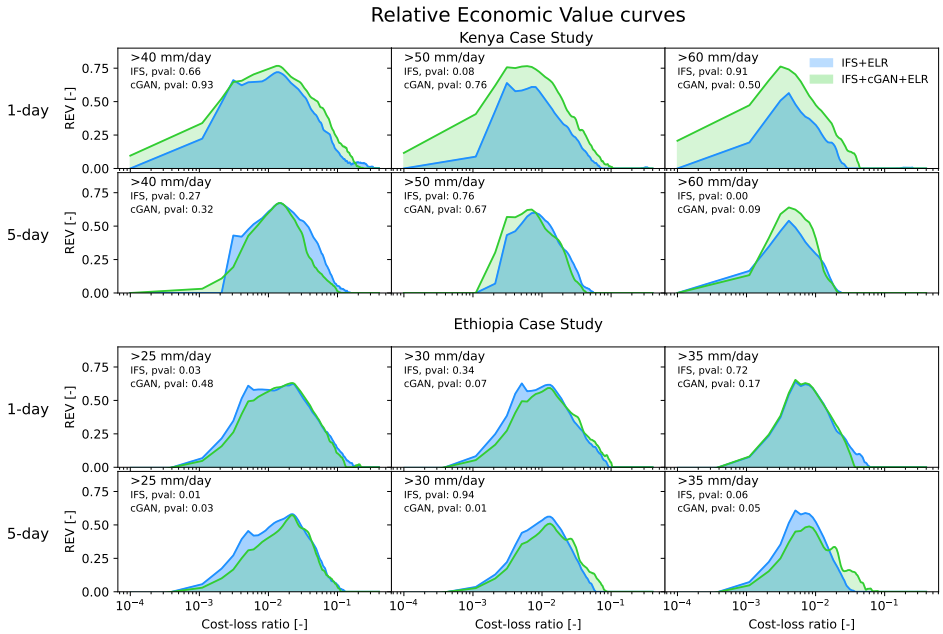


Fig. 5 Relative Economic Value (REV) curves for Kenya (Nairobi, Kiambu and East Kajiado) and Ethiopia (Konso and Moyale) case study regions (refer to Figure A1 for exact locations) at one and five day lead times. A value of one would reflect a perfect forecast that captures all events with no false alarms or misses. Values at zero reflects a forecast that is as good as a default of always or never forecasting the event. P-values summarise the significance of the area under the REV curves, with a lower p-value indicating lower robustness against the sample set used during calculation. Note that the names when displaying the p-value are shortened due to space such that "IFS" refers to IFS+ELR and "cGAN" refers to IFS+cGAN+ELR.

Within the Kenyan case-study region, IFS+cGAN+ELR yields larger REV values than IFS+ELR across all cost-loss ratios and lead times, with the exception of the >40 mm/day threshold at five-day lead times. Higher rainfall thresholds show larger improvements, consistent with the improvement in cross-entropy loss seen in Figure C3. In contrast, the Ethiopian case-study region, shows minimal improvement in REV from the IFS+cGAN+ELR at one-day lead times, with the greatest gains emerging at five-day lead times, particularly at higher cost-loss ratios (>0.3), as seen for the 30 mm/day and 35 mm/day thresholds. While the IFS+cGAN+ELR system demonstrates notable improvement in cross-entropy loss over IFS+ELR at five-day lead times (see Figure C3), this suggests that most of this improvement arises from specific probability ranges. This underscores the importance of considering multiple evaluation metrics when identifying both the rainfall thresholds and the probability ranges within which the greatest forecast value can be realised.

As a final step, we demonstrate the possibility of calibrating forecasts towards rain-gauge observations over Nairobi, Kenya where rain-gauge data was evaluated in-house alongsidelocal forecasters (as described in Section 5.1). The REV curves obtained for

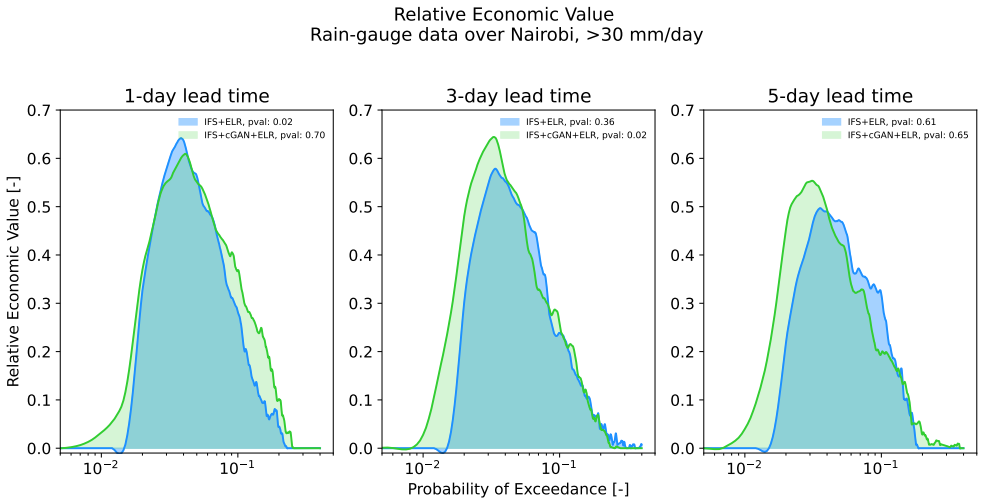


Fig. 6 Relative Economic Value (REV) curves for rain-gauge data, Nairobi at >30 mm/day and one, three and five day lead times. P-values summarise the significance of the area under the REV curves, with a lower p-value indicating lower robustness against the sample set used during calculation.

IFS+ELR and IFS+cGAN+ELR are shown in Figure 6 for the >30 mm/day threshold at one-, three-, and five-day lead times. For supporting information on the relative changes in skill from ELR-based calibration using rain-gauge data, refer to Figure C4.

At one-day lead times, IFS+cGAN+ELR shows the greatest improvement over IFS+ELR at cost-loss ratios below 0.01 and above 0.07 — that is, for decision-makers whose action threshold lies either well below or well above the event's climatological frequency. In contrast, IFS+ELR displays higher REV at cost-loss ratios approaching the climatological event frequency itself (0.02–0.04). At three- and five-day lead times, the advantage of IFS+cGAN+ELR narrows but persists at the lowest cost-loss ratios (below 0.01). Low cost-loss ratios represent decision-makers for whom the cost of acting is small relative to the potential loss from an unmitigated event, and who can therefore tolerate more false alarms to rarely miss a true event; high cost-loss ratios represent the converse, where action is costly and greater confidence is required before acting. This indicates that the cGAN better partitions the relative observed frequency of events across a wider range of predicted probabilities than the IFS alone, appealing to a wider range of risk tolerances and better distinguishing event occurrence away from the climatological frequency – indicating improved forecast resolution following Murphy's decomposition [26].

Peak REV values for IFS+ELR decay quicker with increasing lead time than those for IFS+cGAN+ELR, suggesting that the cGAN-based system retains greater forecast value at longer lead times. This is of particular significance for Anticipatory Action, where extended lead times are critical for enabling timely preparedness and response. The slower degradation of REV in the IFS+cGAN+ELR system implies that decision-makers can rely on it over a broader range of cost-loss ratios and at longer horizons,

widening the window within which actionable and economically valuable warnings can be issued.

4 Discussion

AI-based post-processing methods have emerged as promising and scalable methods to provide high-resolution forecasts tailored to regional needs and priorities. Through low-cost ensemble generation, such systems can improve risk assessments of weather extremes that are crucial for Early Warning Systems and Anticipatory Action. In this study, we evaluated a generative AI method that post-processes operational outputs from ECMWF's Integrated Forecasting System (IFS) outputs with a conditional Generative Adversarial Network (cGAN) to produce probabilistic rainfall forecasts over East Africa, a region with high vulnerability to rainfall extremes. The system has been routinely used in operational centres within East Africa since 2023, providing a unique test case of the utility of AI-based forecasts for Early Warning Systems, particularly under resource-constrained settings.

Using an Ensemble Logistic Regression (ELR) uncertainty quantification approach, we calibrated forecasted distributions to individual probabilities of exceedances for user-defined rainfall thresholds, providing specialised extreme event forecasts. Applying the ELR to both post-processed (IFS+cGAN+ELR) and raw (IFS+ELR) forecasts enabled evaluation of the forecast's reliability [22, 25] at user-relevant rainfall thresholds and for each rainfall threshold further assessment of the forecast's Relative Economic Value (REV) [27, 28] across varying probabilities/cost-loss ratios. Additionally, benchmarking against Isotonic Distributional Regression applied to the IFS (IFS+IDR) provided insight into the relative merits of threshold-specific versus full-distribution calibration approaches.

Regional analysis shows that the threshold-specific ELR calibration consistently yields the largest skill gains at extreme thresholds, highlighting the value of specialised threshold-calibrated forecasts. This is particularly relevant for Early Warning Systems and Anticipatory Action, where decisions to trigger protective measures are typically tied to specific rainfall thresholds that carry known consequences. By calibrating directly at these operationally meaningful thresholds, ELR provides probability estimates tailored to the exceedance levels that decision-makers care about, rather than relying on improvements distributed across the full rainfall distribution that may not translate to the critical tails. Interestingly, IFS+IDR remains competitive with IFS+cGAN across most regions, suggesting that classical statistical post-processing can correct a substantial portion of the calibration error. Nonetheless, the combination of IFS+cGAN+ELR provides the largest cumulative improvement, as the IFS+cGAN provides a stronger baseline upon which ELR can build. Category of reliability analysis further highlights more consistent performance of the IFS+cGAN+ELR system compared to IFS+ELR, with "useful" to "marginally useful" forecasts across most climate zones. A notable improvement in arid and semi-arid areas extending over Somalia, Northern Kenya and North Sudan at five-day lead times is also observed, where the IFS+cGAN+ELR systems retain utility whereas IFS+ELR degrades to having no resolution and therefore being "not useful".

The Relative Economic Value (REV) framework quantifies forecast value for a decision-maker choosing whether to act, based on the cost of acting relative to the loss incurred if an event occurs and no action is taken. For a given rainfall threshold, this cost-loss ratio corresponds directly to a probability of exceedance: low cost-loss ratios suit decision-makers for whom the cost of action is negligible relative to the potential loss, and who are therefore willing to act at low probabilities of exceedance, while high cost-loss ratios suit those who require greater confidence and therefore act only at higher probabilities of exceedance. By considering the forecast Relative Economic Value (REV) across different cost-loss ratios (equivalent to probabilities), we are able to further isolate the probability ranges within which the most skill gain is possible for select case study regions and rain-gauge data. These probability ranges reflect rainfall threshold-specific risk levels within which the most value can be achieved by using the ELR-based uncertainty quantification on the post-processed IFS rather than the raw IFS forecasts. Across both case-study regions and rain-gauges, results indicate that the added value of the calibrated post-processed forecasts is not always uniform but can be concentrated within specific probability ranges that vary by region, rainfall threshold, and lead time. This reinforces the importance of moving beyond aggregate skill metrics towards evaluation frameworks that directly map forecast quality onto the decision space of end-users. In agreement with the improved reliability at longer lead times seen in the regional analysis, the rain-gauge REV analysis shows peak REV values decaying quicker with increasing lead time for IFS+ELR than for IFS+cGAN+ELR. This demonstrates that applying uncertainty quantification to the post-processed forecasts can improve overall forecast value at longer horizons — a property of particular significance for Early Warning Systems and Anticipatory Action, where extended lead times are critical for enabling timely preparedness and response. Together, these findings provide a framework for end-users to identify the rainfall thresholds and probability ranges within which the most actionable skill can be gained, extending the evaluation beyond traditional forecast verification towards decision-relevant assessment.

In future, there are still some improvements that could be made to the ELR uncertainty quantification step. For example, as the ELR is fitted for individual thresholds, the predicted probabilities between thresholds may not be mutually consistent. Possible additions could therefore include Extended Logistic Regression approaches [29], which further include the predictand threshold as an input parameter, or more precise distributional regression methods that directly predict rainfall distribution parameters [30]. In using the evaluation results to appropriately make decisions, there is also further work to be done. On one hand, an end-user could know which cost-loss ratios are most relevant to them for a given rainfall threshold and individually assess the added value of the forecast. However, for usual forecasters it is difficult to know the cost-loss ratios applicable to all end-users that they cater to. Moreover in issuing warnings, context such as number of previous heavy rain days would vastly affect the risk associated with a future heavy rain event. As a first step, we have made all evaluation results interactively available alongside the forecasts (see python package here: https://github.com/Fenwick-Cooper/SEWAA-forecasts). This could help better guide decisions of issuing alerts under different contexts and across a wider range of end-users.

Despite these caveats, the combination of low-cost large ensembles, uncertainty quantification, and decision-centric evaluation offers a practical route to make AI post-processing systems operationally useful for Early Warning Systems, especially where computational resources are constrained. We anticipate the framework will transfer to other hazards and regions with modest re-training, provided suitable truth data and representative cost–loss contexts are available.

5 Methods

5.1 Data

Numerical Weather Prediction (NWP) operational rainfall forecasts of the Integrated Forecast System (IFS) from the European Centre for Medium-Range Weather Forecasting (ECMWF) are used [17] and are obtained at a 6-hourly resolution on a $0.1^\circ \times 0.1^\circ$ grid. cGAN forecasts are obtained at a 24-hourly resolution at one to five day lead times, again on a $0.1^\circ \times 0.1^\circ$ grid. All 6-hourly forecasts are processed to yield 24-hour accumulations. 24-hour accumulations are obtained starting at 6AM for the lead time of interest till 6AM of the next day (i.e., for a one day lead time, rainfall is accumulated from 30 to 54 hour lead times and so on). This is to maintain consistency with the accumulation times of point rain-gauge data as measured by the local meteorological services.

For evaluation, we use the IMERGv7 satellite rainfall product [31] as the ground truth data. This is also the target data used in training the cGAN as detailed in subsection 5.4. IMERGv7 data is obtained at a 30-minute resolution on a $0.1^\circ \times 0.1^\circ$ grid and is accumulated to obtain 24-hour accumulated rainfall. For additional evaluation, we use 24-hour accumulated rainfall data measured at six rain-gauge stations over the Nairobi province in Kenya, namely: Jomo Kenyatta International Airport, Wilson Airport, Eastleigh, Dagoretti, Kabete and Ngong.

In our analysis, we maintain a train-test split of train years 2018, 2019 and 2020 and test years 2021, 2023 and 2024. Where available, we use all days within the train and test years. However, due to data limitations, we only use months going from June-December and March-June for test years 2023 and 2024 respectively.

5.2 ELR: an interpretable equivalent to a Rectified Linear Unit (ReLU) Neural Network

We are interested in obtaining the probability of exceedance for a user-defined threshold, x, from an ensemble rainfall forecast consisting of a non-decreasing sequence of n members $\mathbf{X}^t = X_1^t, X_2^t, \ldots, X_n^t$ with lead-time t,

$$P(X > x) = f(\mathbf{X}^t) \tag{1}$$

We thus employ an uncertainty quantification step that further calibrates the forecasted distribution to an individual probability of exceedance and is similar to more recent AI fine-tuning methods. A typical, AI-based fine-tuning would apply a ReLU Neural Network with one hidden layer of r neurons and one linear output layer, represented as follows,

$$\psi^{ReLu(r)}(\mathbf{X^t}) = \sum_{i=1}^{r} \beta_i \cdot max\{\mathbf{W}_i \cdot \mathbf{X^t} + b_i, 0\} \tag{2}$$

with β_i, b_i and $\mathbf{W}_i \in \mathbb{R}^{1\times n}$ being trainable coefficients. To then obtain probabilities of exceedances, a sigmoid layer would be applied,

$$\sigma(\psi^{ReLu(r)}) = \frac{1}{1 + e^{-\psi^{ReLu(r)}}} \tag{3}$$

Under certain assumptions (i.e., uniform initialisation of coefficient weights), this is equivalent to well-established methods of Kernel Logistic Regression (KLR) [32] where a non-linear transformation is applied to the inputs before applying the logistic regression [33, 34].

For our modelling problem, each ensemble member represents a discrete sample from the true underlying forecasted distribution. We utilise this knowledge to apply a transformation that approximates the true underlying distribution before fitting the logistic regression. We fit a linear, B(asis)-spline [35, 36] – detailed in sub-section 5.3 – as an equivalent variable kernel method [37] to $\mathbf{X}^t$, which approximates $max\{\mathbf{W}_i \cdot \mathbf{X^t} + b_i, 0\}$ with knot points at i. Applying a logistic regression thereafter leads us to Equation 3.

In summary, ELR maps an ensemble forecast to a corresponding probability of threshold exceedance. It does so by approximating the underlying forecast distribution from which to obtain well-calibrated probabilities of exceedances, presenting an interpretable equivalent to a single ReLU Neural Network. It can be trained to represent grid-cell level data or even point, gauge data located on the forecast grid as shown in Figure 1.

5.3 B-spline Fitting and Hyperparameter Search

Second-order linear B-splines are fitted following the Cox-de Boor algorithm [35]. Accordingly, second-order B-splines are recursively composed of first-order B-splines which are described as,

$$B_{i,1}(X) = \begin{cases} 1 & \text{if } X_i \leq X < X_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where X_i and X_{i+1} are a non-decreasing sequence denoting the knot interval for knot i. The function $B_{i,1}$ is right-continuous and respects the partition of unity, i.e.,

$$\sum_i B_{i,1}(X) = 1, \quad \text{for all X} \tag{5}$$

The second-order B-splines are obtained from $B_{i,1}$ as,

$$B_{i,2} = \frac{X - X_i}{X_{i+1} - X_i} \cdot B_{i,1} + \frac{X_{i+1} - X}{X_{i+2} - X_{i+1}} \cdot B_{i+1,1} \tag{6}$$

Considering the recursive composition of $B_{i,2}$ from $B_{i,1}$ – which respects the partition of unity property – $B_{i,2}$ is bounded within the interval $[X_i, X_{i+2}]$ such that each spline basis is a local function.

In fitting the B-spline to approximate the underlying ensemble distribution, the non-decreasing sequence of the ensemble rainfall forecast, $\mathbf{X}^t$ is decomposed into its second-order B-spline components at each member $X_1^t, \ldots, X_n^t$. The number of knot points was specified to be between 5, 10, or 15, and was chosen using K-fold cross-validation, with K=5. Spacing of the knot interval X_i and X_{i+1} was further cross-validated for between the options of uniformly- or quantile- spaced intervals.

5.4 cGAN architecture and training

We consider an AI post-processing approach described by [18], which consists of a cGAN trained using the Wasserstein loss function [38]. The cGAN architecture is composed of residual blocks [39] that allow for stable optimisation with deeper networks. Such an architecture has also been demonstrated for rainfall prediction over Switzerland and the U.K. [15, 16]. The cGAN takes 14 forecast variables (listed in Table D1) from IFS, obtained at a 6-hourly temporal and $0.1^\circ \times 0.1^\circ$ spatial resolution, as inputs (see [18, 19] for further details), alongside elevation and land-sea mask values. It is trained to 24-hourly rainfall forecasts at one to five day lead times, at a $0.1^\circ \times 0.1^\circ$ spatial resolution. Years 2018, 2019 and 2020 are used as training years and 2021 is used as the validation set.

5.5 Reliability Analysis and Cost-Loss Frameworks

The ELR is used to calibrate raw (IFS+ELR) and post-processed (IFS+cGAN+ELR) IFS towards user-relevant thresholds. As a reference, we compare to uncalibrated IFS and IFS+cGAN ensemble outputs, alongside Isotonic Distributional Regression based calibration of IFS (IFS+IDR) [23]. We assess the reliability and resolution of probabilities obtained following the Consistency, Optimality and Reproducibility (CORP) approach [22]. The CORP method leverages monotonicity assumptions within the reliability diagram – that considers observed frequencies as a function of predicted probabilities, where perfect reliability indicates a one-to-one relationship – to non-parametrically bin data using the Pool Adjacent Violators (PAV) algorithm. Given that the set of extreme rainfall events is sparse however, obtaining an unbiased estimate of calibration error over a finite sample size is difficult [40, 41]. To mitigate this, we apply a kernel smoothing on the space of observed events before computing the PAV-based bins [42–44]. From the reliability diagrams obtained, we empirically calculate the reliability error (REL) – i.e., distance from the perfect reliability line (lower is better) – and resolution (RES) – i.e., distance from a climatological prediction (higher is better) – terms, alongside their surrounding uncertainties. These are then used to categorise the utility of forecasts in a similar manner to [25]. Accordingly, having REL significantly less than RES is "useful"; for REL greater than or equal to RES we further consider having RES greater than zero as "marginally useful" – as the forecasts are still better than a mean climatological prediction – and RES significantly equal to zero are classified as "not useful".

Cost-loss frameworks [27, 28] are further used to calculate the REV of forecasts across a range of cost of action (C) to loss of inaction (L) ratios,

$$REV(\frac{C}{L}) = \frac{E_{fcst} - E_{clim}}{E_{perfect} - E_{clim}} \tag{7}$$

Where the economic value of a forecast (E_{fcst}) is expressed in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), which sake of simplicity are taken as fractions normalised by the total number of events,

$$E_{fcst} = (TP + FP) \cdot C + FN \cdot L \tag{8}$$

The economic value of a climatological forecast is the minimum between either always, or never acting on average,

$$E_{clim} = min((TP + FN) \cdot L, C) \tag{9}$$

And the economic value of a perfect forecast is,

$$E_{perfect} = (TP + FN) \cdot C \tag{10}$$

When the forecasted probability of an event, p, is well-calibrated, one can set $\frac{C}{L} = p$. Calculating the REV as a function of ELR predicted probabilities thus allows us to scan the probability space and define areas with highest added economic value. The REV is sensitive to the sample set over which it is calculated, with some events having a disproportionate effect on its value. To estimate the robustness of the REV curves obtained, we therefore consider the area under the curve as a summary statistic and compute it for 100 bootstrap resampled sets of data as well as for the full dataset without resampling. We then compute a one-sample t-test to evaluate how well the area under the curve obtained without resampling matches its distribution obtained from bootstrap resampling. All p-values are reported alongside their respective REV curves.

Supplementary information. No Supplementary Information

Acknowledgements. The authors would like to acknowledge funding from Google.org through the United Nations World Food Programme.

Conflict of interest. The author's declare no conflict of interest.

Code and Data Availability. All code used is readily available on GitHub at https://github.com/Fenwick-Cooper/SEWAA-forecasts. IFS data used is publicly available through the ECMWF MARS server.

Author Contributions. SN conceptualised and designed the study with assistance from DK, DM and FC. SN and DK conducted preliminary implementation of the ELR using data generated by FC. Analysis, evaluation and writing was done by SN with guidance from DK, DM, FC, TP, AW and AT. MAm, MAr, IO and JM helped in bringing perspective of the practical utility towards Anticipatory Action. All authors contributed to reviewing and streamlining the paper.

Appendix A Map of selected case study regions

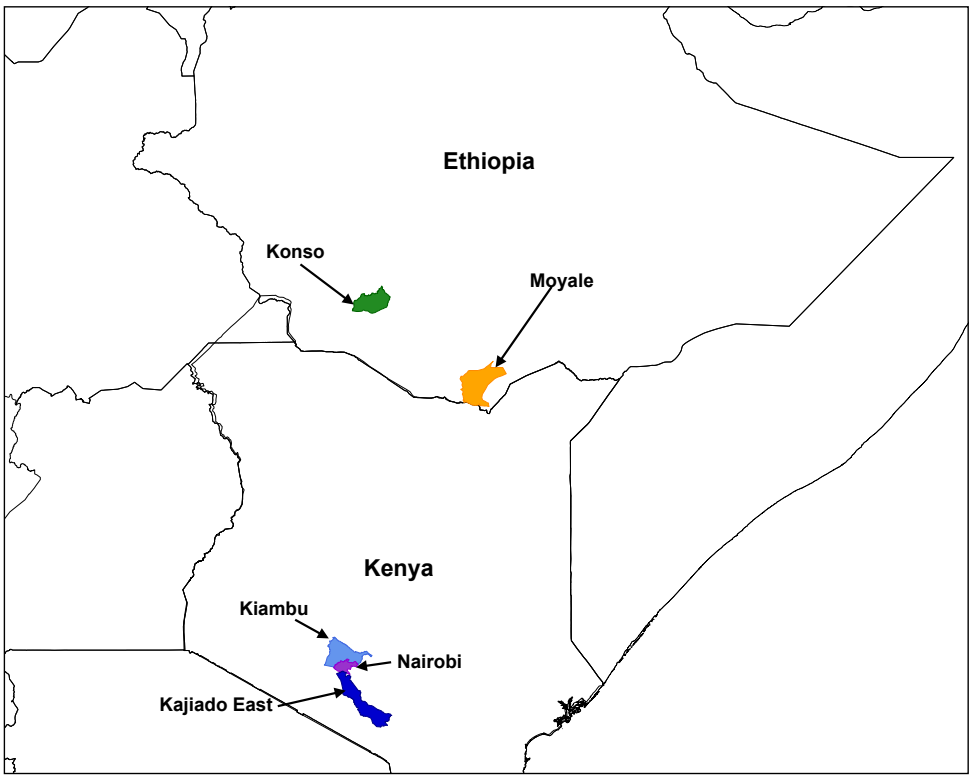


Fig. A1 Locations of case studies used in Section 2. Note that the border between Kenya, South Sudan and Ethiopia is ambiguously demarcated, representing the Ilemi Triangle which is disputed territory.

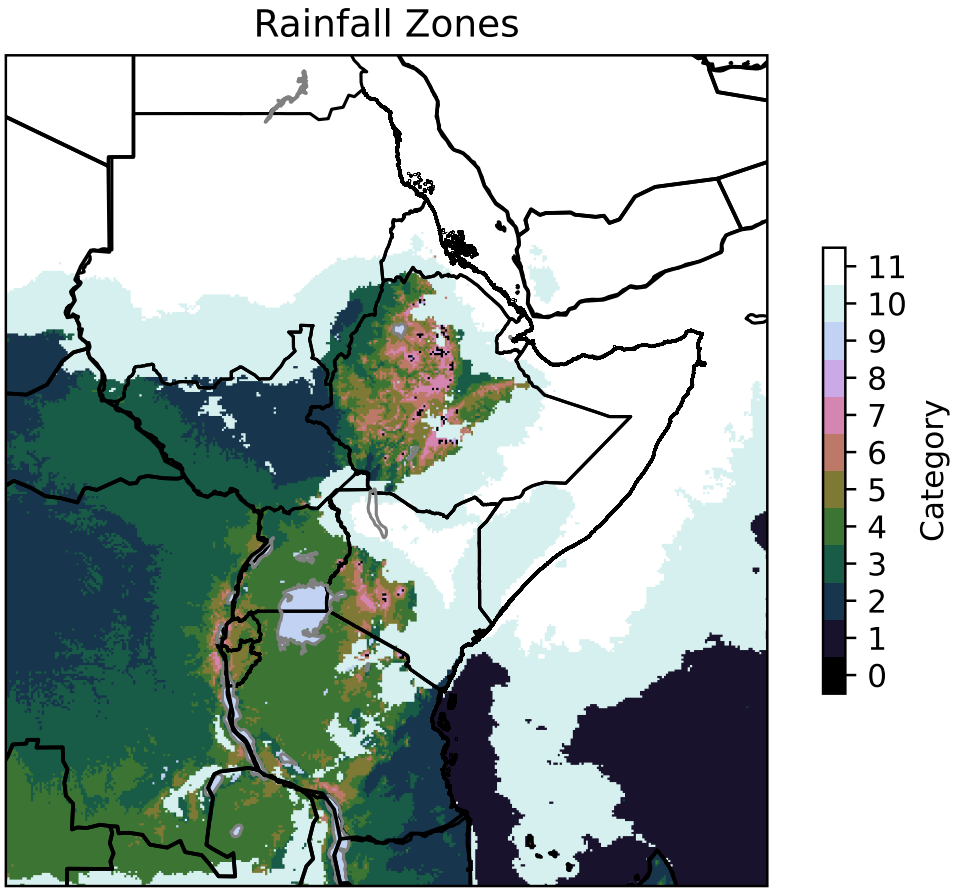


Fig. B2 Sub-regional rainfall zones selected in a similar fashion to Köppen-Geiger climate zones [24] using criteria of elevation, lakes and total annual rainfall received in years 2018-2021. An additional criteria that each zone contains a sufficient number of samples such that drier regions encompass larger areas to compensate for inadequate rainfall was also used.

Appendix B Rainfall Zones

Appendix C Improvement from ELR-based output for case-study and rain-gauge data

Appendix D List of variables used within the cGAN

References

[1] Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters,

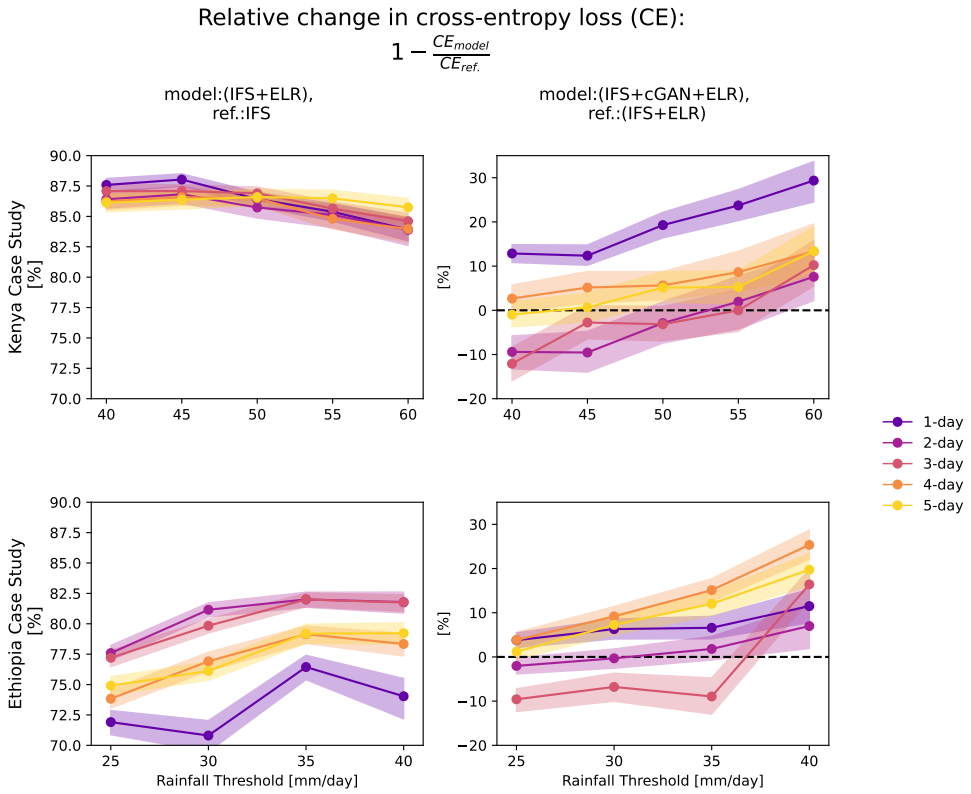


Fig. C3 Relative change in cross-entropy loss (or CE in short) for the Kenya (Nairobi, Kiambu and East Kajiado), top panel, and Ethiopia (Konso and Moyale), bottom panel, case study regions. First column shows the relative changes from IFS+ELR versus empirical ensemble-based probabilities of IFS. Second column shows the relative change from IFS+cGAN+ELR forecasts with reference to IFS+ELR forecasts. Relative changes are plotted in percentages and a higher value is better. Uncertainty ranges indicate the 95% confidence interval as obtained from bootstrap resampling done 1000 times.

D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M.: Probabilistic weather forecasting with machine learning. Nature **637**, 84–90 (2025) https://doi.org/10.1038/s41586-024-08252-9

[2] Bouallègue, Z.B., Clare, M.C.A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J.S., Lang, S.T.K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., Pappenberger, F.: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. Bulletin of the American Meteorological Society **105**, 864–883 (2024) https://doi.org/10.1175/BAMS-D-23-0162.1

[3] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range

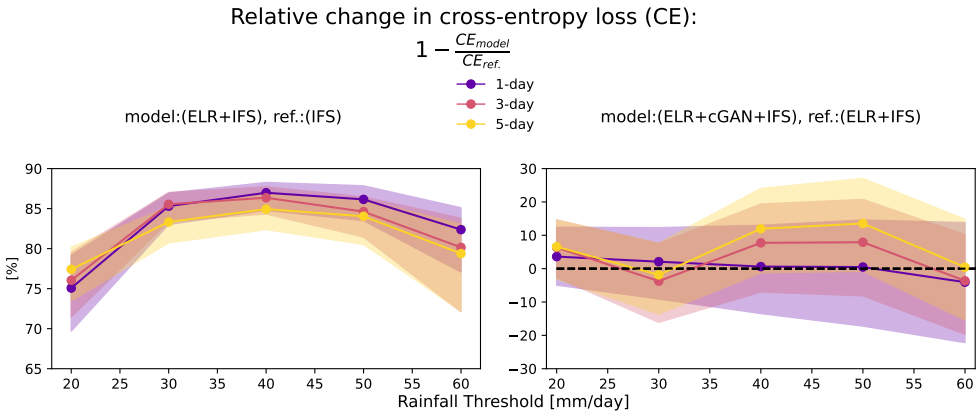


Fig. C4 Similar to Figure 2 but for rain-gauge data over Nairobi and only the ELR-based 1000-ensemble member cGAN output is shown.

Table D1 Inputs to the cGAN model sourced from the IFS forecast model.

ECMWF code	Predictor
CP	Convective precipitation
TP	Total precipitation. (Used as the IFS prediction of rainfall.)
MCC	Medium cloud cover. (Cloud cover at medium altitude.)
SP	Surface pressure
SSR	Surface incoming solar radiation
T2M	Two metre temperature
TCIW	Total column ice water. (The total ice water in a column of air within the grid box and between the surface and the top of the atmosphere.)
TCLW	Total column liquid water
TCRW	Total column rain water
TCW	Total column water
TCWV	Total column water vapour
U (700 hPa)	Zonal (South to North) wind at 700 hPa
V (700 hPa)	Meridional (West to East) wind at 700 hPa
Elevation	Nearest neighbour interpolated to the IMERG grid.
Land-sea-mask	Average fraction of water, including ocean lakes and rivers, within an IMERG grid cell.

global weather forecasting with 3d neural networks. Nature **619**, 533–538 (2023) https://doi.org/10.1038/s41586-023-06185-3

[4] Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., Li, H.: Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. npj Climate and Atmospheric Science **6**, 190 (2023) https://doi.org/10.1038/s41612-023-00512-1

[5] Wainwright, C.M., Finney, D.L., Kilavi, M., Black, E., Marsham, J.H.: Extreme rainfall in east africa, october 2019–january 2020 and context under future climate change. Weather **76**, 26–31 (2021) https://doi.org/10.1002/wea.3824

[6] Chang'a, L.B., Kijazi, A.L., Mafuru, K.B., Kondowe, A.L., Osima, S.E., Mtongori, H.I., Ng'ongolo, H.K., Juma, O.H., Michael, E.: Assessment of the evolution and socio-economic impacts of extreme rainfall events in october 2019 over the east africa. Atmospheric and Climate Sciences **10**, 319–338 (2020) https://doi.org/10.4236/acs.2020.103018

[7] (WMO), W.M.O.: Early Warning Systems for all plan unveiled at COP30 (2022). https://wmo.int/media/news/early-warnings-all-action-plan-unveiled-cop27

[8] Gudoshava, M., Otieno, G., Koech, E., Misiani, H., Ongoma, J.G., Heinrich-Mertsching, C., Wachana, C., Endris, H.S., Mwanthi, A., Kilavi, M., Mwangi, E., Colman, A., Parker, D., Mutemi, J.N., Machio, P., Omay, P.O., Ombai, P., Anande, D., Kondowe, A., Mugume, I., Ayabagabo, P., Houssein, H.Y., Waiss, M.S., Abeshu, B., Kayoya, E., Sharawe, M.N., Bahaga, T., Todd, M., Segele, Z., Atheru, Z., Artan, G.: Advances, gaps and way forward in provision of climate services over the greater horn of africa. Frontiers in Climate **6** (2024) https://doi.org/10.3389/fclim.2024.1307535

[9] Fink, A.H., Engel, T., Ermert, V., Van Der Linden, R., Schneidewind, M., Redl, R., Afiesimama, E., Thiaw, W.M., Yorke, C., Evans, M., et al.: Mean climate and seasonal cycle. Meteorology of tropical West Africa: The forecasters' handbook, 1–39 (2017)

[10] Mwanthi, A.M., Mutemi, J.N., Opijah, F.J., Mutua, F.M., Atheru, Z., Artan, G.: Implications of wrf model resolutions on resolving rainfall variability with topography over east africa. Frontiers in Climate **6** (2024) https://doi.org/10.3389/fclim.2024.1311088

[11] Satheesh, A.R., Knippertz, P., Fink, A.H.: Machine learning models for daily rainfall forecasting in northern tropical africa using tropical wave predictors. Weather and Forecasting **40**, 1895–1916 (2025) https://doi.org/10.1175/WAF-D-24-0192.1

[12] Yuval, J., Langmore, I., Kochkov, D., Hoyer, S.: Neural general circulation models optimized to predict satellite-based precipitation observations (2024). https://arxiv.org/abs/2412.11973

[13] Scheuerer, M., Heinrich-Mertsching, C., Bahaga, T.K., Gudoshava, M., Thorarinsdottir, T.L.: Applications of machine learning to predict seasonal precipitation for East Africa (2024). https://arxiv.org/abs/2409.06238

[14] Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M.A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J.,

Pritchard, M.: Residual corrective diffusion modeling for km-scale atmospheric downscaling. Communications Earth & Environment **6**, 124 (2025) https://doi.org/10.1038/s43247-025-02042-5

[15] Leinonen, J., Nerini, D., Berne, A.: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing **59**(9), 7211–7223 (2021) https://doi.org/10.1109/TGRS.2020.3032790

[16] Harris, L., McRae, A.T.T., Chantry, M., Dueben, P.D., Palmer, T.N.: A generative deep learning approach to stochastic downscaling of precipitation forecasts. Journal of Advances in Modeling Earth Systems **14**(10) (2022) https://doi.org/10.1029/2022MS003120

[17] ECMWF: Ifs documentation, part v: Ensemble prediction system. Technical report, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK (November 2020)

[18] Antonio, B., McRae, A.T.T., MacLeod, D., Cooper, F.C., Marsham, J., Aitchison, L., Palmer, T.N., Watson, P.A.G.: Postprocessing east african rainfall forecasts using a generative machine learning model. Journal of Advances in Modeling Earth Systems **17**(3) (2025) https://doi.org/10.1029/2024MS004796

[19] Cooper, F.C., Nath, S., McRae, A.T.T., Antonio, B., Weisheimer, A., Palmer, T., Gudoshava, M., Kalladath, N., Amidhun, A., Kinyua, J., Kimani, H., Koros, D., Mwai, Z., Maswi, C., Chanzu, B., Abebe, S., Tamene, B., Kebebe, B., Teshome, A., Pappenberger, F., Chantry, M., Obai, I., Mason, J.: Rainfall forecasts in daily use over East Africa improved by machine learning (2025). https://arxiv.org/abs/2512.24525

[20] Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. Technical report (2000)

[21] Richardson, D.S.: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quarterly Journal of the Royal Meteorological Society **127**(577), 2473–2489 (2001) https://doi.org/10.1002/qj.49712757715

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

[23] Walz, E.-M., Henzi, A., Ziegel, J., Gneiting, T.: Easy uncertainty quantification (easyuq): Generating predictive distributions from single-valued model output. SIAM Rev. **66**, 91–122 (2024)

[24] Geiger, R.: Klassifikation der klimate nach w. köppen. Landolt-Börnstein – Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik **3**, 603–607 (1954)

[25] Weisheimer, A., Palmer, T.N.: On the reliability of seasonal climate forecasts. Journal of The Royal Society Interface **11**(96), 20131162 (2014) https://doi.org/10.1098/rsif.2013.1162

[26] Murphy, A.H.: A new vector partition of the probability score. Journal of Applied Meteorology **12**, 595–600 (1973)

[27] Wilks, D.S.: A skill score based on economic value for probability forecasts. Meteorological Applications **8**(2), 209–219 (2001) https://doi.org/10.1017/S1350482701002092

[28] Palmer, T.N.: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Quarterly Journal of the Royal Meteorological Society **128**(581), 747–774 (2002) https://doi.org/10.1256/0035900021643593

[29] Wilks, D.S.: Extending logistic regression to provide full-probability-distribution mos forecasts. Meteorological Applications **16**(3), 361–368 (2009) https://doi.org/10.1002/met.134

[30] Bülte, C., Horat, N., Quinting, J., Lerch, S.: Uncertainty quantification for data-driven weather models. Artificial Intelligence for the Earth Systems **5**(1), 240049 (2026) https://doi.org/10.1175/AIES-D-24-0049.1

[31] Huffman, G.J., Bolvin, D.T., Joyce, R., Kelley, O.A., Nelkin, E.J., Tan, J., Watters, D.C., West, B.J.: Integrated multi-satellite retrievals for gpm (imerg) technical documentation. Technical report (2023)

[32] Green, P.J., Silverman, B.W.: Nonparametric Regression and Generalized Linear Models. Chapman and Hall/CRC, ??? (1993). https://doi.org/10.1201/b15710

[33] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks (2020)

[34] Guyomard, M., Barbosa, S., Fillatre, L.: Kernel logistic regression approximation of an understandable relu neural network. In: Proceedings of the 40th International Conference on Machine Learning, pp. 12268–12291 (2023)

[35] Boor, C.: On calculating with b-splines. Journal of Approximation theory **6**(1), 50–62 (1972)

[36] Boor, C.: Splines as linear combinations of b-splines,. Approximation Theory II, 1–47 (1976)

[37] Silverman, B.W.: Spline smoothing: The equivalent variable kernel method.

Technical Report 3 (1984)

[38] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 214–223 (2017). https://doi.org/10.48550/arXiv.1701.07875

[39] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[40] Roelofs, R., Cain, N., Shlens, J., Mozer, M.C.: Mitigating bias in calibration error estimation. In: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (eds.) Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 4036–4054. PMLR, ??? (2022). https://proceedings.mlr.press/v151/roelofs22a.html

[41] Błasiok, J., Gopalan, P., Hu, L., Nakkiran, P.: A unifying theory of distance from calibration. In: Proceedings of the 55th Annual ACM Symposium on Theory of Computing. STOC 2023, pp. 1727–1740. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3564246.3585108 . https://doi.org/10.1145/3564246.3585108

[42] Bröcker, J.: Some remarks on the reliability of categorical probability forecasts. Monthly Weather Review **136**(11), 4488–4502 (2008) https://doi.org/10.1175/2008MWR2329.1

[43] Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., Xu, C.: Metrics of calibration for probabilistic predictions (2022). https://arxiv.org/abs/2205.09680

[44] Błasiok, J., Nakkiran, P.: Smooth ece: Principled reliability diagrams via kernel smoothing. In: The Twelfth International Conference on Learning Representations (ICLR) (2024). https://openreview.net/forum?id=9G3z6A1X5r