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, hybrid physical and AI systems

Shruti Nath

shruti.nath@physics.ox.ac.uk

University of Oxford https://orcid.org/0009-0007-4449-1621

David Koros

Kenya Meteorological Department

Fenwick Cooper

University of Oxford

David MacLeod

University of Cardiff

Hannah Kimani

Kenya Meteorological Department

Zacharia Mwai

Kenya Meteorological Department

Christine Maswi

Kenya Meteorological Department

Bernard Chanzu

Kenya Meteorological Department

Asaminew Teshome

Ethiopian Meteorological Institute

Bekalu Tamene

Ethiopian Meteorological Institute

Bekele Kebebe

Ethiopian Meteorological Institute

Samrawit Abebe

Ethiopian Meteorological Institute

Masilin Gudoshava

Climate Prediction and Application Centre, Inter-Governmental Authority on Development (ICPAC).

Ahmed Amdihun

Climate Prediction and Application Centre, Inter-Governmental Authority on Development (ICPAC).

Isaac Obai

United Nations, World Food Programme

Maurine Ambani

United Nations, World Food Programme
Mark Arango
United Nations, World Food Programme
Jesse Mason
United Nations, World Food Programme
Matthew Chantry
ECMWF https://orcid.org/0000-0002-1132-0961
Florian Pappenberger
ECMWF https://orcid.org/0000-0003-1766-2898
Antje Weisheimer
University of Oxford https://orcid.org/0000-0002-7231-6974
Tim Palmer
University of Oxford https://orcid.org/0000-0002-7121-2196

Article

Keywords: Rainfall, ML/AI, extremes, forecast value, Early Warning Systems

Posted Date: January 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8534913/v1

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

Additional Declarations: There is **NO** Competing Interest.

Predicting East Africa's rainfall extremes with calibrated, hybrid physical and AI 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.
[2]Kenya Meteorological Department.
[3]School of Earth and Environmental Sciences, University of Cardiff.
[4]Ethiopian Meteorological Institute.
[5]Climate Prediction and Application Centre, Inter-Governmental Authority on Development (ICPAC).
[6]UN WFP.
[7]European Centre for Medium-range Weather Forecasting, U.K..
[8]National Centre for Atmospheric Sciences, 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). Hybrid Artificial Intelligence (AI) approaches, which integrate global physical forecasts with AI, 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, fine-tuning method that calibrates ensemble forecasts to probabilities of exceedances for rainfall extremes. Applied

to physical and hybrid AI rainfall forecasts over East Africa, case-study results show encouraging skill improvement from hybrid AI systems at higher rainfall thresholds. Sub-regional analysis shows the hybrid AI systems better reflecting the observed chances of extremes at longer lead times. Forecast skill decomposed across rainfall and probability thresholds further 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, forecast 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. Nevertheless, 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 contributions come from deep convective systems [8, 9], resolution is 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 that combine physical models with AI are termed "hybrid" approaches and 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 a hybrid AI rainfall forecasting system prototyped within Kenyan and Ethiopian meteorological centres

alongside the Greater Horn of Africa's Climate Prediction and Application Centre (ICPAC) since 2023. The hybrid AI system consists of a conditional Generative Adversarial Network (cGAN) that post-processes global, operational 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 system routinely used by regional and local meteorological centres in East Africa for real-time post-processing of global forecasts to produce corrected, probabilistic rainfall forecasts. This presents a unique testbed to demonstrate the practical utility of hybrid AI forecasting 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 a fine-tuning step that calibrates the hybrid AI system from focussing on the full rainfall distribution, to specialising on individual user-defined extreme rainfall thresholds. The fine-tuning is easily applied to raw IFS data as well, allowing for fair assessment of the added value from the hybrid AI system.

This study introduces an Ensemble Logistic Regression (ELR) fine-tuning step depicted in Figure 1 and described in Section 5.2, 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 equivalent to traditional AI fine-tuning approaches and calibrates to probabilities without compressing the ensemble to a mean/variance summary. Across two case-study regions (Nairobi–Kiambu–East Kajiado, Kenya; Konso–Moyale, Ethiopia), sub-regional maps, and Nairobi rain-gauges, we assess how well forecasted probabilities of individual rainfall thresholds match their actual observed frequencies, as derived from reliability-diagram slope values [22]. At each rainfall threshold, we further assess the forecast Relative Economic Value (REV) across various cost–loss ratios (analogous to risk levels) [23, 24], to identify ranges of risk within which the AI-enhanced forecasts can add most value. By combining threshold-wise calibration with decision-oriented value metrics, our framework identifies when and where AI-enhanced forecasts are most useful for Early Warning Systems. We further make the evaluation results available alongside the forecasts to guide warning issuance under varying risk contexts.

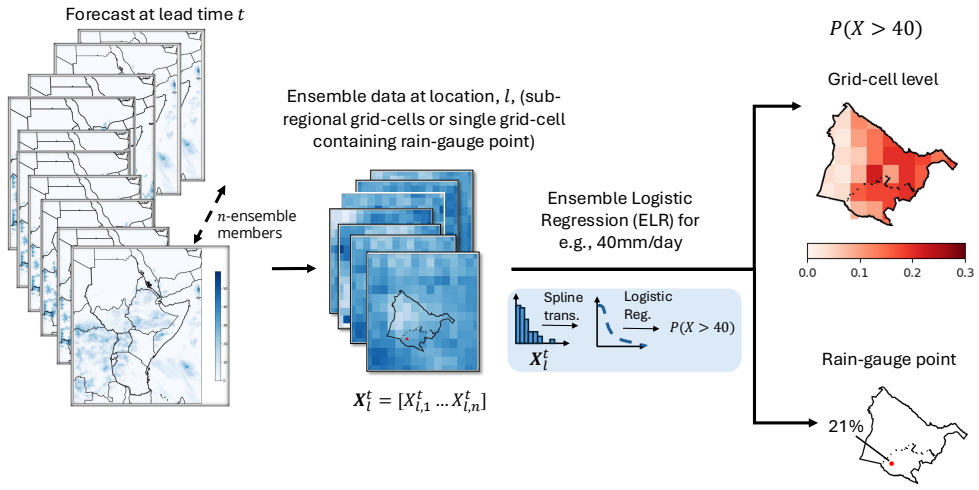


Fig. 1 Framework for fine-tuning ensemble output with Ensemble Logistic Regression (ELR). ELR is trained on specific sub-regions within the forecast domain. It takes ensemble, grid-cell level data as input and can be trained to predict probabilities of exceedance for a given threshold (40 mm/day shown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

2 Case-study showcase

We initially demonstrate and evaluate the ELR-based fine-tuning in predicting grid-cell level probabilities of exceedance for 24-hour accumulated rainfall thresholds at two select case-study regions. A separate ELR is fitted per region by pooling all grid points within that region. We consider the performance of ELR-tuned IFS forecasts alongside that of ELR-tuned hybrid, cGAN and IFS forecasts. When training the ELR, 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 for both the ELR and the cGAN. Details of the IFS, cGAN and truth rainfall datasets used can be found in Section 5.1.

Regions encompassing Nairobi, Kiambu and East Kajiado in Kenya, and Konso and Moyale in Ethiopia were chosen in collaboration with the local meteorological departments (see Figure A1 for details), representing areas that faced significant floods and flash floods during the test period. The region in Kenya is central and mountainous, typically receiving rain of >30mm/day during the rainy seasons, while the regions in Ethiopia are within the drier, lowland region and are prone to flash floods at rainfall levels as little as 20 mm/day. We therefore examine rainfall thresholds of 40-60 mm/day and 25-40 mm/day at 5 mm/day steps for the Kenya and Ethiopia case-study regions respectively.

Figure 2 shows the relative change in skill from ELR-based fine-tuning. The cross-entropy loss, which is the log-likelihood of a Bernoulli distribution and the typical

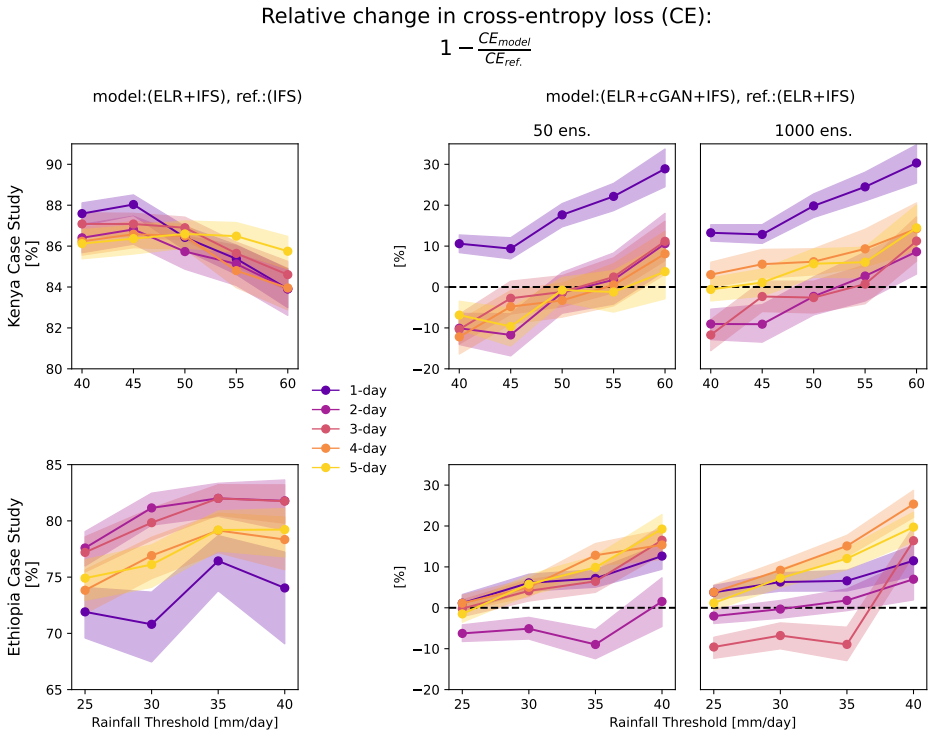


Fig. 2 Relative change in cross-entropy loss (or CE in short) for the Kenya (Nairobi, Kiambu and East Kajiado), top panel, and Ethiopia (Konso and Moyale), bottom panel, case study regions. First column shows the relative changes from fine-tuned ELR-based versus empirical ensemble-based probabilities of IFS. Second and third columns show the relative change from ELR-based fine-tuned cGAN+IFS 50- and 1000- member ensemble forecasts respectively with reference to ELR-based fine-tuned IFS 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.

optimisation target for classification tasks, is used as a measure of skill. We consider the relative change in skill obtained from ELR-tuned IFS forecasts with reference to empirical IFS ensemble probabilities, alongside the additional changes when using ELR-tuned 50- and 1000- ensemble member hybrid AI, cGAN+IFS (referred to as just "cGAN" for simplicity) forecasts with reference to ELR-tuned IFS forecasts. Across all thresholds and lead times, the ELR-tuned IFS displays a 70-90 % improvement relative to empirical IFS probabilities, showing a clear benefit from the ELR-based calibration. At lower thresholds of less than 50 mm/day and 35 mm/day for the Kenya and Ethiopia case study regions respectively, the ELR-tuned cGAN displays limited improvement of at most 20 % relative to the ELR-tuned IFS and confined to specific lead times. Particularly for the Kenya case study region, the ELR-tuned 50-ensemble cGAN shows a degradation in skill at thresholds lower than 50 mm/day and lead-times greater than two days, while the ELR-tuned 1000-ensemble cGAN displays a

similar degradation but only at two and three day lead times. Interestingly, a similar degradation in skill at two and three day lead times is seen for lower thresholds within the Ethiopia case study region as well, suggesting limited skill of the ELR-tuned cGAN within these lead-time windows. For higher thresholds, the ELR-tuned cGAN shows larger improvements relative to the ELR-tuned IFS across all lead-times. Especially at four to five days lead-time, these improvements are more pronounced for the 1000- as compared to 50- ensemble cGAN, which is as expected since larger ensembles should sample the forecast distribution better and be less overconfident [21].

To further investigate the ELR-tuned forecast reliability across rainfall thresholds, we consider the slope values of the reliability diagram that measure how closely forecasted probabilities match observed frequencies of an event [22] (see Section 5.5 for details on calculation). A slope of one indicates perfect reliability, while slopes below 0.5 suggest little or no skill added to the Brier Skill Score in reference to climatology [25]. Slope values that are above one indicate that for each unit change in predicted probabilities, the change in actual observed frequencies is larger suggesting underconfidence and vice versa for slope values below one. For simplicity, we focus on the cGAN 1000-member ensemble here on after to demonstrate the additional operational gain from low-cost, large ensemble generation. Results from the reliability analysis are shown in Figure 3.

For the Kenya case study region, both the cGAN and IFS systems exhibit slope values greater than 0.5 across all rainfall thresholds and lead times, indicating forecast skill beyond just climatology. Across all rainfall thresholds at one, four and five day lead times, cGAN slope values remain above one, indicating a tendency towards under-confidence in forecasts, while at two and three day lead times cGAN slope values remain equal to or below one, indicating a tendency towards overconfidence in forecasts. In contrast, IFS displays median slope values larger than one, with a decline towards the no skill line within range of the error bars with increasing rainfall thresholds. This indicates that the poorer skill seen for ELR-tuned cGAN relative to ELR-tuned IFS in Figure 2 at two to three day lead times could arise from overconfidence in forecasted probabilities.

In the Ethiopia case study region, cGAN and IFS show similar reliability at one to three day lead times, with a decline towards no skill with increasing rainfall thresholds. Especially at three day lead times, the cGAN displays no skill for 35 mm/day which is again in line with the poorer skill seen in Figure 2. However, at four and five day lead times, the cGAN demonstrates improved reliability with slope values close to one, whereas IFS remains near no skill.

To assess the probability ranges within each rainfall threshold where the most added value from the forecasts can be gained, we consider the forecasts' Relative Economic Value (REV) [23, 24]. At each probability of threshold exceedance value (interchangeable as cost-loss ratio or risk level), the REV places the forecast on a scale from a default forecast that always or never predicts the event (zero on the scale) to an ideal forecast that perfectly captures all events with no false alarms or misses (one on the scale). Figure 4 provides REV values obtained for different rainfall thresholds. For simplicity, we focus on one and five day lead times representative of the shortest and longest lead times predicted by the cGAN. Within the Kenyan case study region,

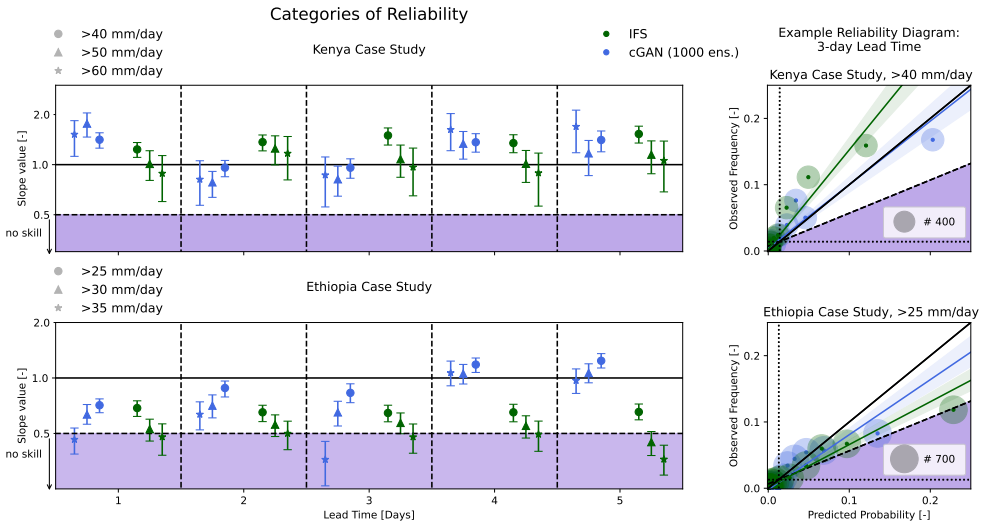


Fig. 3 Reliability analysis conducted across lead times and rainfall thresholds for IFS (green) and cGAN (blue). Slope coefficients of the reliability diagram are considered for Kenya (Nairobi, Kiambu and East Kajiado) and Ethiopia (Konso and Moyale) case study regions. A value close to one reflects perfect reliability whereas a value below 0.5 (purple shaded area) reflects little or no skill beyond climatology. Uncertainty ranges indicate the 95% confidence interval as obtained from bootstrap resampling done 1000 times. Example reliability diagrams are shown in the right insets, depicting the number of samples within each probability bin (area of the circle) and the corresponding reliability slopes fitted. Perfect reliability (solid) and no-skill (dashed) lines are given as reference, with the region below the no-skill line shaded purple indicating slope values below 0.5.

the cGAN displays larger REV values as compared to IFS across all cost-loss ratios and lead times, with the exception of >40 mm/day at five day lead times. Higher rainfall thresholds show larger improvements, which is consistent with the results from Figure 2. The Ethiopia case study region shows minimal improvement in REV from cGAN at one day lead times and most at five day lead times, particularly at higher cost-loss ratios of > 0.3, as seen for 30 mm/day and 35 mm/day thresholds. This is consistent with the improved reliability seen at four and five day lead times in Figure 3 and suggests that most of the skill gain seen in Figure 2 comes from select probability thresholds, emphasising the need to consider several metrics when identifying both rainfall thresholds and probability ranges within which the most value can be achieved from a forecast.

3 Regional and rain-gauge analysis

To demonstrate the fine-tuned skill over the whole East Africa region, a separate ELR-tuning is done over grid-cell level data for each sub-region. Sub-regions are chosen in a similar fashion to Köppen-Geiger climate zones [26] 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

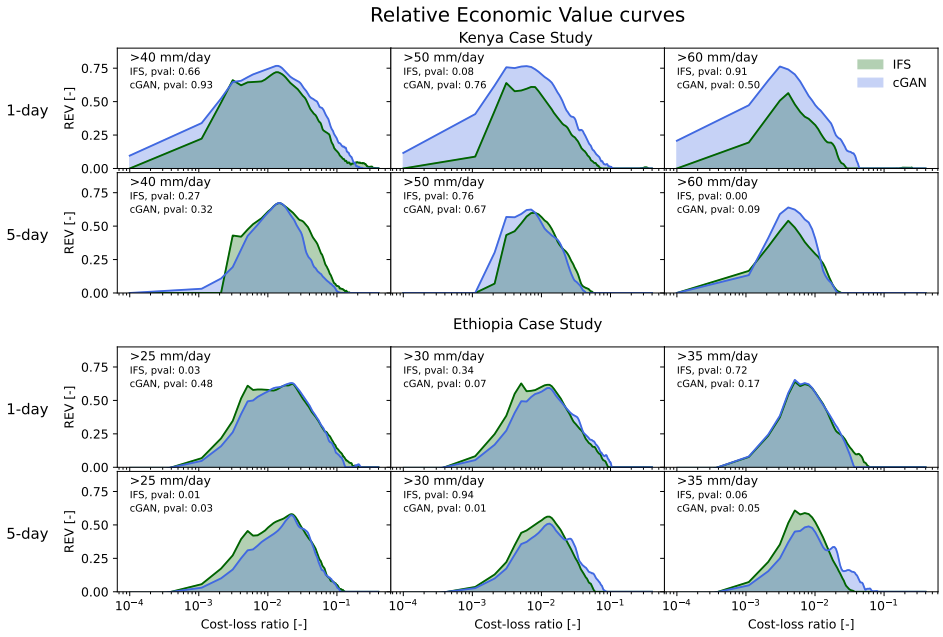


Fig. 4 Relative Economic Value (REV) curves for Kenya (Nairobi, Kiambu and East Kajiado) and Ethiopia (Konso and Moyale) case study regions at one and five day lead times. A value of one would reflect a perfect forecast that captures all events with no false alarms or misses. Values at zero reflects 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.

larger areas to compensate for inadequate rainfall (see Figure C3). Rainfall thresholds of 20-60 mm/day at 10 mm/day steps are fine-tuned for. Categories of reliability are calculated by pooling together all grid-cells in a given sub-region as shown in Figure 5. A category of "perfect" indicates reliability diagram slope values near one, whereas "still useful" and "marginally useful" indicate slope values between one and 0.5 and less than 0.5 respectively. Reliability diagram slope values near zero are assigned the category of "not useful" and negative slope values are assigned the category of "dangerous". The highest rainfall threshold up to which at least one of the forecasting systems has "perfect" to "still useful" reliability is shown (unless this does not exist, in which case "not useful" to "marginally useful" is shown instead). It should be noted that the hatched area is particularly dry and its threshold is therefore fixed at >5 mm/day. Supporting information on the sub-regional skill gained through ELR-based fine-tuning is provided in Figure D4

Panel (b) of Figure 5 summarizes the number of sub-regions falling within each category of reliability for the cGAN and IFS at one, three and five day lead times, and shows a clear shift from "useful" to "marginally/not useful" with longer lead times for both systems. Despite this, the cGAN retains more regions classified as "perfect" or "still useful" across all lead times compared to IFS, demonstrating improved forecast

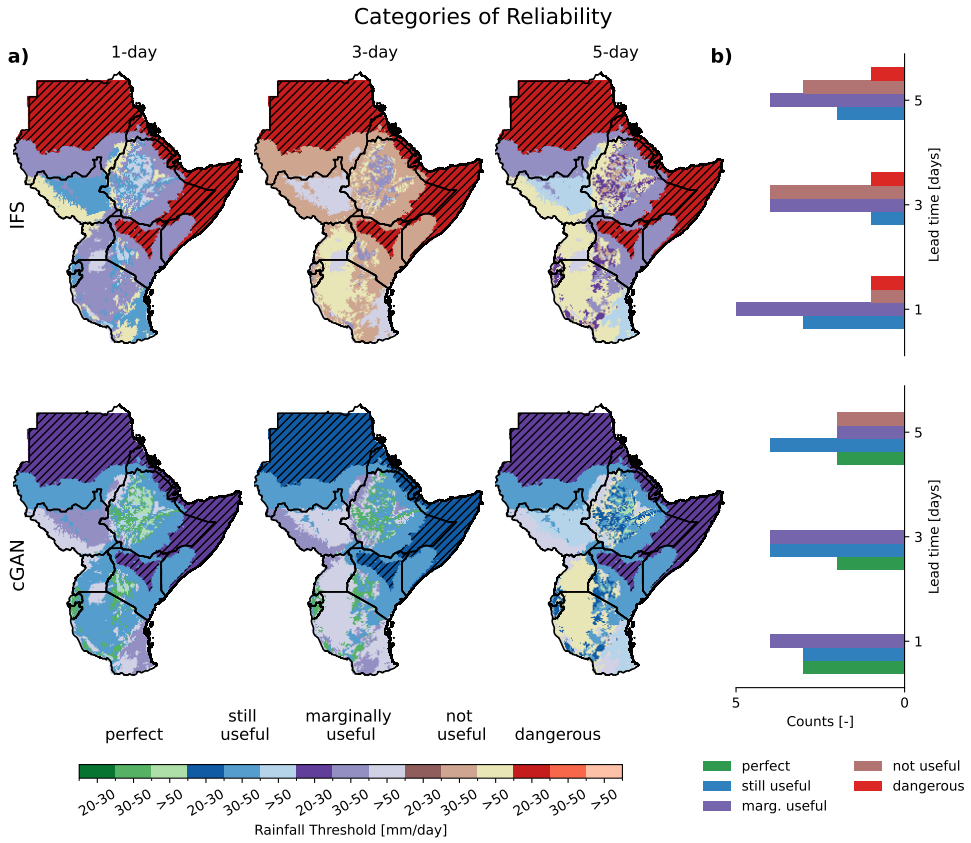


Fig. 5 Categories of reliability from ELR-based fine-tuning of IFS and cGAN at one, three and five day lead times. For each region and lead time, the highest rainfall threshold for which at least one of the forecasting systems has a useful to perfect forecast is shown (unless this does not exist, in which case useless to marginally useful is shown instead). Panel (a) provides maps of categories of reliability according to rainfall zones (see Figure C3) and hatched area represents a significantly dry region where rainfall threshold of >5 mm/day is fixed. Panel (b) summarizes the maps into counts of the number of rainfall zones within each category of reliability.

reliability. Panel (a) shows the geographical distribution of reliability, highlighting mountainous regions of Ethiopian Highlands, Mount Kenya, Mount Kilimanjaro, and the Rwenzori range, where the cGAN maintains "perfect" to "still useful" reliability across all lead times. Conversely, areas around Lake Victoria in Uganda and North Tanzania show "not useful" reliability across both systems at five day lead times, indicating limited predictability for these regions at longer lead times.

As a final step, we demonstrate the possibility to fine-tune towards rain-gauge data over Nairobi (as described in section 5.1). The reliability diagram and REV obtained for cGAN and IFS are shown in Figure 6 for >30 mm/day at one and five days lead time. For supporting information on the relative changes in skill from ELR-based fine-tuning using rain-gauge data, refer to Figure D5. At both one and five day lead times, reliability diagrams depict near perfect reliability for both IFS and cGAN, although cGAN shows a slightly poorer reliability with an offset towards under-forecasting the event. When considering REV, cGAN shows most improvement on top of IFS at cost-loss ratios less than 0.01 and more than 0.07, while IFS displays higher REV at cost-loss ratios approaching the event climatological frequency of $0.02-0.04$. This indicates that the cGAN better partitions the relative observed frequency of events across a wider range of predicted probabilities as compared to IFS. As such, it can better distinguish the occurrence of events at probabilities away from the climatological frequency of the event, indicating better forecast resolution following Murphy's decomposition [27].

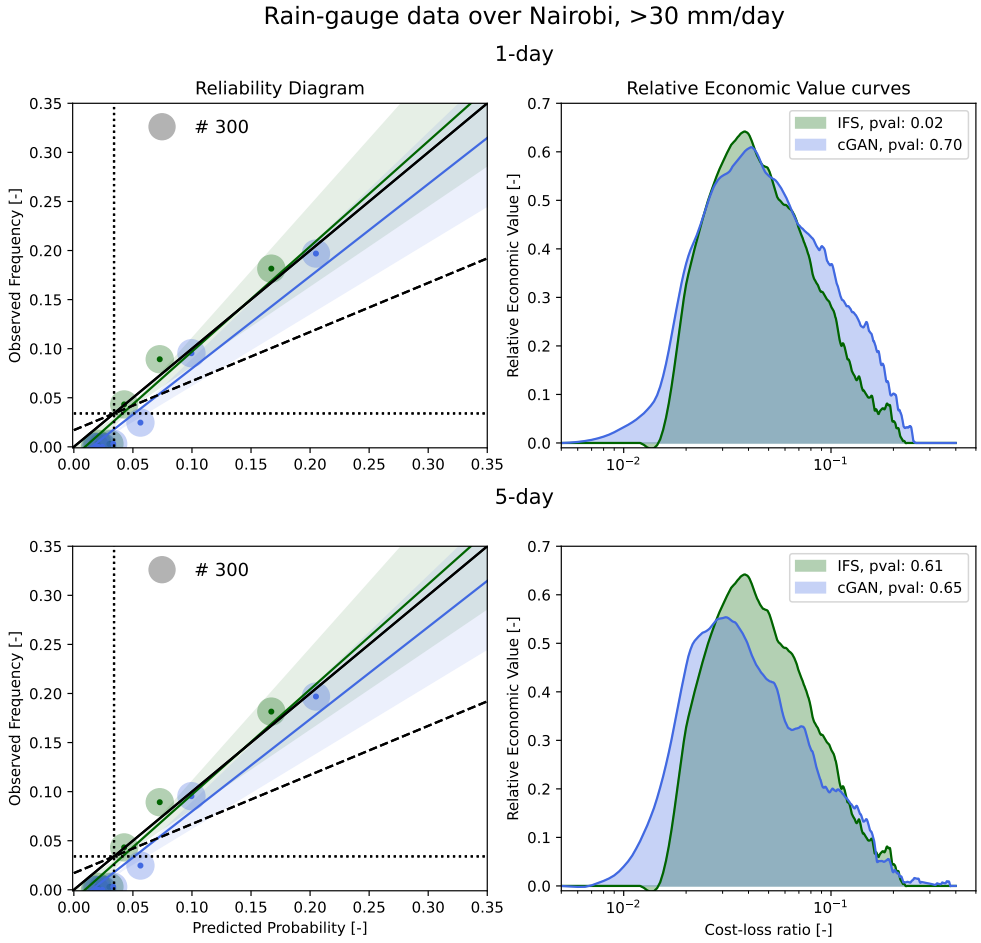


Fig. 6 Reliability diagrams and Relative Economic Value (REV) curves for rain-gauge data, Nairobi at >30 mm/day and one and five day lead times. Uncertainty ranges within the reliability diagram indicate the 95 % confidence interval as obtained from bootstrap resampling done 1000 times. P-values summarise the significance of the area under the REV curves, with a lower p-value indicating lower robustness against the sample set used during calculation.

4 Discussion

Hybrid physical and AI forecasting systems 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 hybrid AI system that post-processes operational outputs from the 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) fine-tuning approach, we calibrated forecasted distributions to individual probabilities of exceedance for user-defined rainfall thresholds, providing specialised extreme event forecasts. Applying the ELR to both hybrid AI (cGAN+IFS) and traditional physical (IFS) models enabled evaluation of the forecast's reliability [22] at user-relevant rainfall thresholds and for each rainfall threshold further assessment of the forecast's Relative Economic Value (REV) [23, 24] across varying probabilities/cost-loss ratios.

Case-study and sub-regional reliability analysis demonstrates comparable if not improved reliability from the hybrid AI system relative to IFS, with predicted probabilities better matching the observed frequency of events, especially at higher rainfall thresholds and longer lead times. For example, in the Ethiopia case study region at four and five day lead times, the hybrid AI system demonstrates reliability diagram slope values close to one – meaning a near perfect match between predicted probabilities and observed frequencies of an event – whereas those of IFS approach no added skill with reference to a climatological forecast [25]. By considering the 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 from the hybrid AI system. These probability ranges reflect rainfall threshold-specific risk levels within which the most value can be achieved by using the hybrid AI system as compared to IFS. This provides a framework for end-users to identify rainfall thresholds and probability ranges within which most actionable skill can be gained, giving results that extend beyond traditional forecast evaluation to decision-making.

In future, there are still some improvements that could be made to the ELR fine-tuning 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 [28], which further include the predictand threshold as an input parameter, or the use of monotonicity constraints (e.g., isotonic regression) to ensure consistency across thresholds. 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, interpretable fine-tuning, and decision-centric evaluation offers a practical route to make hybrid AI forecasts 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 $0.1° \times 0.1°$ grid. cGAN forecasts are obtained at a 24-hourly resolution at one to five day lead times, again on a $0.1° \times 0.1°$ 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 [29] 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 $0.1° \times 0.1°$ 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 a fine-tuning step that further calibrates the forecasted distribution to an individual probability of exceedance. 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) [30] where a non-linear transformation is applied to the inputs before applying the logistic regression [31, 32].

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 [33, 34] – detailed in sub-section 5.3 – as an equivalent variable kernel method [35] 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. For further consideration on ELR in comparison to more traditional Ensemble Model Output Statistic (EMOS) methods, refer to Appendix B.

5.3 B-spline Fitting and Hyperparameter Search

Second-order linear B-splines are fitted following the Cox-de Boor algorithm [33]. 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.,

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

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 a hybrid AI forecasting system described by [18], which consists of a cGAN trained using the Wasserstein loss function [36]. The cGAN architecture is composed of residual blocks [37] 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 E1) 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 Categories of Reliability and Cost-Loss Frameworks

We assess the reliability of ELR-based probabilities obtained from fine-tuning both IFS- and cGAN- ensemble forecasts following [22]. ELR-based predicted probabilities are binned and the fraction of positives (i.e., observed frequency of event) within each bin calculated. Bins are chosen by selecting quantiles of the predicted probabilities going from zero to one with a step size of 0.1. The process is repeated on 1000 bootstrap resampled sets of data. A weighted linear regression using the number of samples at each bin as weights is done for each bootstrap, to obtain 1000 slope values. A slope value significantly equal to one indicates a one-to-one relationship between predicted probability and observed frequency, and thus a perfect forecast. Consequently, assessing the slope value and uncertainty range allows categorisation of the forecast reliability into "perfect" (slope value near one), "still useful" (slope less than one but at least equal to 0.5), "marginally useful: (slope positive but less than 0.5), "not useful" (slope equal to zero) and "dangerous" (negative slope).

Cost-loss frameworks [23, 24] 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

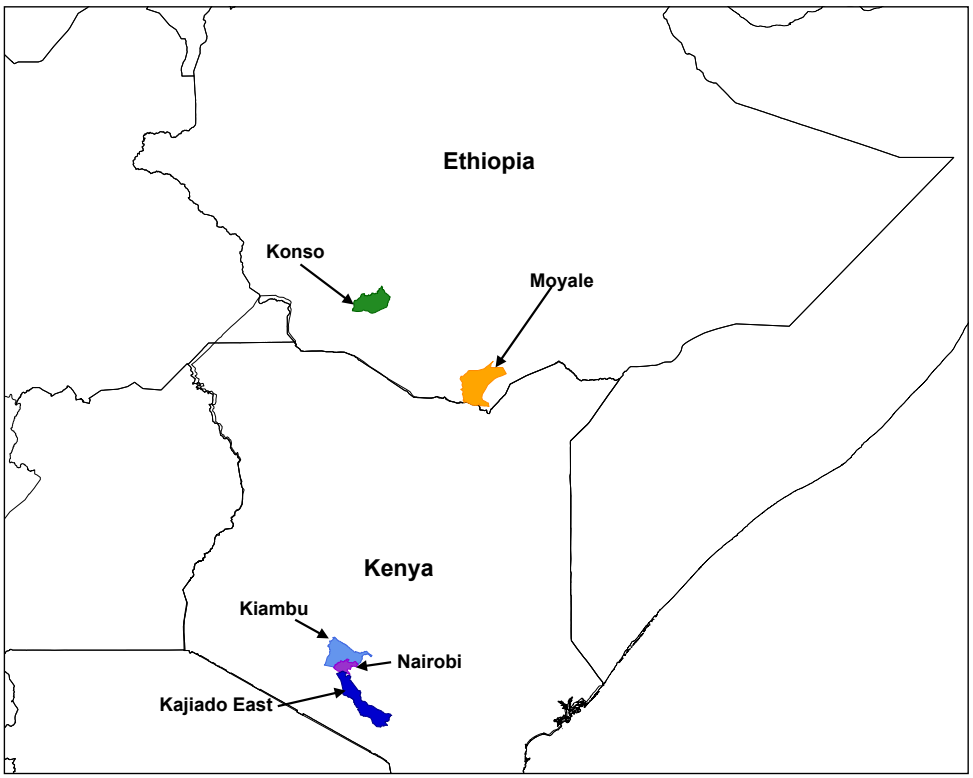


Fig. A1 Locations of case studies used in Section 2. Note that the border between Kenya, South Sudan and Ethiopia is ambiguously demarcated, representing the Ilemi Triangle which is disputed territory.

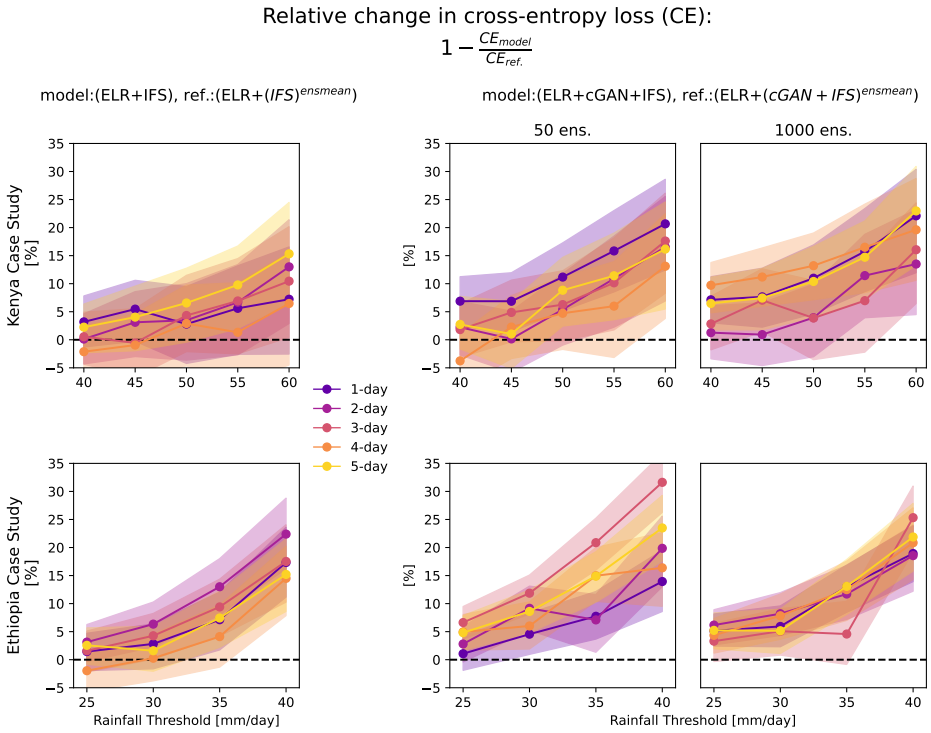


Fig. B2 Relative change in cross-entropy loss when applying ELR to all ensemble members with reference to applying ELR to ensemble mean values for regions Nairobi, Kiambu and North Kajiado (top panel) and Konso and Moyale (bottom panel) and models IFS, cGAN (50 ensemble) and cGAN (1000 ensemble). Relative changes are plotted in percentages and and a higher value is better. Uncertainty ranges indicate the 95% confidence interval as obtained from bootstrap resampling done 1000 times.

Appendix B ELR in comparison to other EMOS methods

ELR is similar to other EMOS methods, although to date, most EMOS methods compress ensemble information into mean and standard deviation metrics (e.g., [38, 39]) before applying logistic regression. We propose applying the non-linear spline transform across all ensemble members (not just ensemble mean), to allow for more efficient extraction of ensemble information before performing the logistic regression. To illustrate this, further tests comparing the performance when using all ensemble members vs. the ensemble mean within the ELR is provided in Figure B2.

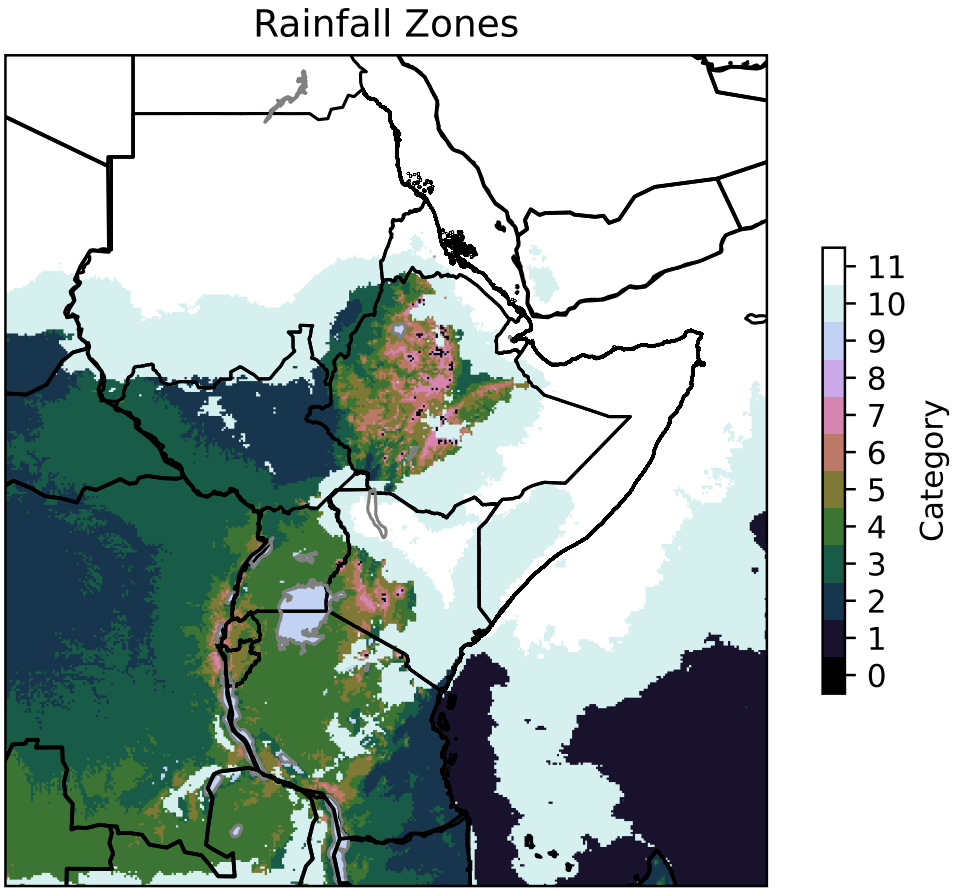


Fig. C3 Sub-regional rainfall zones selected in a similar fashion to Köppen-Geiger climate zones [26] using criteria of elevation, lakes and total annual rainfall received in years 2018-2021. An additional criteria that each zone contains a sufficient number of samples such that drier regions encompass larger areas to compensate for inadequate rainfall was also used.

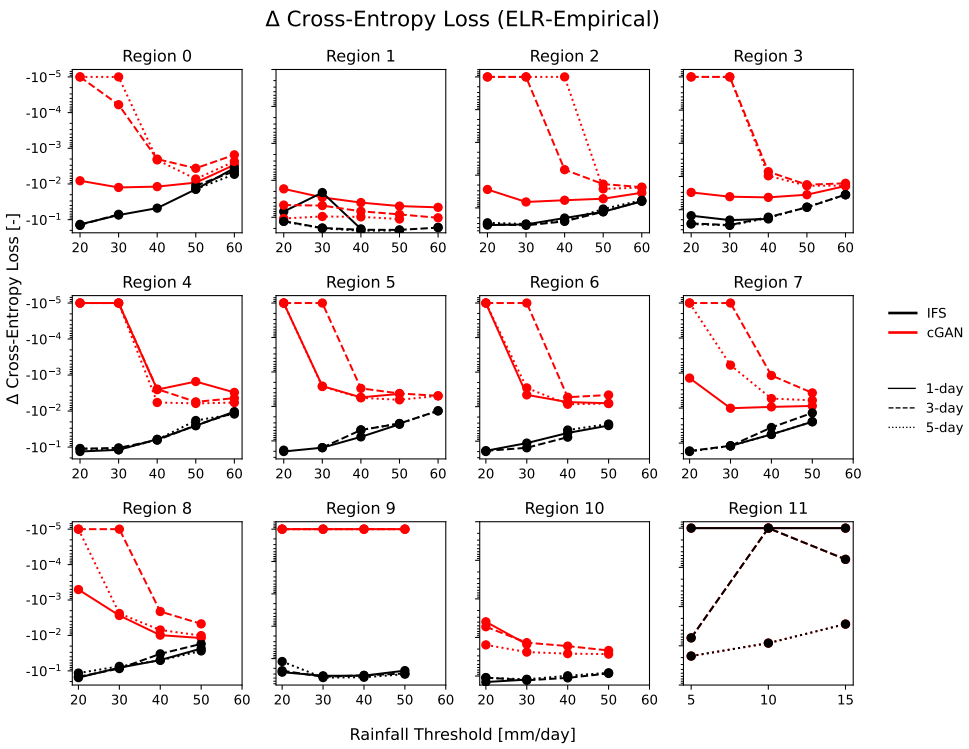


Fig. D4 Change in cross-entropy loss from ELR-based fine-tuning across the 11 rainfall zones shown in Figure C3. Improvement is shown for both cGAN and IFS systems and is as compared to their respective empirical ensemble probabilities.

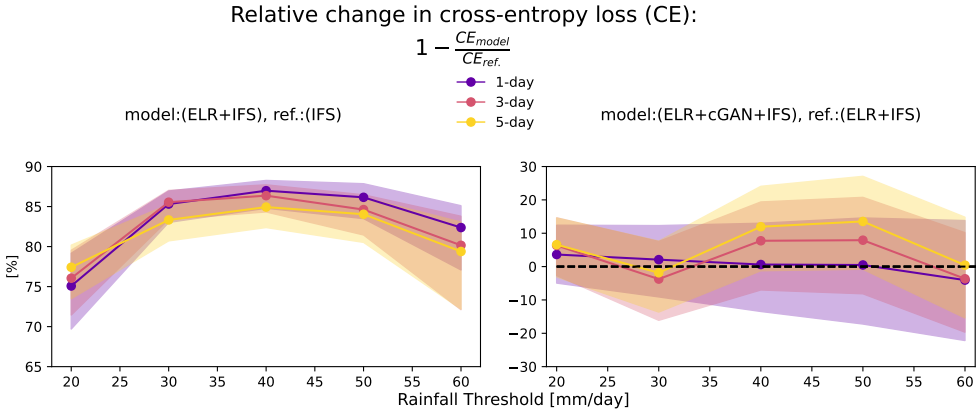


Fig. D5 Similar to Figure 2 but for rain-gauge data over Nairobi and only the ELR-tuned 1000-ensemble member cGAN is shown.

Table E1 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.

Appendix C Rainfall Zones

Appendix D Improvement from ELR-based fine-tuning for regional and rain-gauge data

Appendix E List of variables used within the cGAN

References

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

[23] 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

[24] 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

[25] Mason, S.J.: On using "climatology" as a reference strategy in the brier and ranked probability skill scores. Technical report (2003)

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

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

[28] 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

[29] 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)

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

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

[32] 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)

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

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

[35] Silverman, B.W.: Spline smoothing: The equivalent variable kernel method. Technical Report 3 (1984)

[36] 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

[37] 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

[38] Hamill, T.M., Whitaker, J.S., Wei, X.: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. Monthly Weather Review **132**(6), 1434–1447 (2004) https://doi.org/10.1175/1520-0493(2004)132⟨1434:ERIMFS⟩2.0.CO;2

[39] Sprenger, M., Mahlstein, I., Caratsch, A., Müller, M.: A warning system for the laseyer windstorm in a narrow and deep valley of northeastern switzerland. Weather and Forecasting **39**(10), 1487–1502 (2024) https://doi.org/10.1175/WAF-D-23-0223.1