

Stop using root-mean-square error as a precipitation target!

Kieran M. R. Hunt^{1,2}

¹*Department of Meteorology, University of Reading, Reading, UK*

²*National Centre for Atmospheric Sciences, University of Reading, Reading UK*

September 11, 2025

Abstract

Root-mean-square error (RMSE) remains the default training loss for data-driven precipitation models, despite precipitation being semi-continuous, zero-inflated, strictly non-negative, and heavy-tailed. This Gaussian-implied objective misspecifies the data-generating process because it tolerates negative predictions, underpenalises rare heavy events, and ignores the mass at zero. We suggest replacing RMSE with the Tweedie deviance, a likelihood-based and differentiable loss from the exponential-dispersion family, with variance function $V(\mu) = \mu^p$. For $1 < p < 2$ it yields a compound Poisson–Gamma distribution with a point mass at zero and a continuous density over $y > 0$, matching observed precipitation characteristics. In this paper, we (i) estimate p from the variance–mean power law and show that precipitation across temporal aggregations is far from Gaussian and that the Tweedie power p increases with accumulation length towards a Gamma limit; and (ii) demonstrate consistent skill gains when training deep data-driven models with Tweedie deviance instead of RMSE. In diffusion-model downscaling over Beijing, Tweedie loss improves wet-pixel MAE and extreme recall (~ 0.60 vs 0.50 at the 99th percentile). In ConvLSTM nowcasting over Kolkata, Tweedie loss yields improves wet-pixel MAE and dry-pixel hit rates, with improvements autoregressively compounding over lead time (for MAE, $\sim 2\%$ at $t+1$ growing to $\sim 16\%$ at $t+4$). Because the Tweedie deviance is continuous in p , it adapts smoothly across scales, offering a statistically justified, practical replacement for RMSE in precipitation-based learning tasks.

1 Introduction

1.1 Precipitation distributions

Precipitation is a semi-continuous process. At the temporal scales of interest for weather and hydrology, it exhibits a large point mass at zero, strictly non-negative support, and usually a positive skew with heavy upper tails. These features are inherently non-Gaussian, and a common practical response in the literature is to partition precipitation into an intensity component and an occurrence (or frequency) component. Yet the choice of an appropriate probabilistic model remains contested, with the community having accumulated a remarkably broad set of parametric proposals.

A first class of models treats precipitation as a point process, with storm or cell arrivals modelled by (possibly clustered) Poisson processes. Examples include Poisson or Neyman–Scott formulations for event arrivals and spatiotemporal clustering (e.g., Rodriguez-Iturbe et al., 1987; Cowpertwait, 1994; Cowpertwait et al., 2002; Ramesh et al., 2012) and studies of the extent to which drop counts deviate from Poisson assumptions (e.g., Kostinski and Jameson, 1997). While these approaches make useful stochastic rainfall generators, they do not in themselves prescribe a suitable distribution for nonzero amounts.

A second class of models treats wet period amounts or

time-averaged intensities as approximately Gamma distributed. Foundational statistical approaches (Thom, 1958; Katz, 1977; Wilks, 1990) have led to the Gamma distribution being applied across a range of climates and datasets (e.g., Cho et al., 2004; Segond et al., 2006; Husak et al., 2007; Amburn et al., 2015; Murata et al., 2020). More recently, Martinez-Villalobos and Neelin (2019) gave a physical rationale for the emergence of Gamma-like distributions under temporal averaging of rainfall events.

Many alternatives have been explored beyond the Poisson and Gamma distributions. Mixtures of exponentials have been found to fit nonzero daily amounts better than a single Gamma in certain regions (e.g., Wilks, 1998, 1999; Woolhiser and Roldán, 1982; Rho and Kim, 2019). Lognormal models have also been proposed, particularly for satellite-derived rain rates (e.g., Cho et al., 2004; Gupta and Waymire, 1998). Other families such as Weibull (e.g., Wilks, 1989; Olivera and Heard, 2019) and kappa distributions (e.g., Strong et al., 2025) have also been proposed, and non-parametric or flexible parametric approaches have been explored for cases when a single family seems too restrictive. For extremes, generalised extreme-value (GEV) and generalised Pareto distributions are widely used to represent annual (or other timescale) maxima and threshold exceedances (e.g., Papalexiou and Kout-

soyannis, 2013; AghaKouchak and Nasrollahi, 2010; Phoophiwfa et al., 2024).

Because of the difficulty in asserting a single parametric form for precipitation, many evaluation studies deliberately avoid strong distributional assumptions, instead relying on non-parametric or distribution-agnostic scores. For verification of spatial structure, e.g., for gridded rainfall, neighbourhood or scale-aware methods such as the fractions skill score are common (e.g., Ebert et al., 2004; Roberts and Lean, 2008; Mittermaier, 2021). For event occurrences, categorical scores (e.g., the Brier score) are popular (e.g., Brier et al., 1950; Hogan et al., 2010). For continuous probabilistic assessments, proper scores such as the continuous ranked probability score (CRPS) and feature-based metrics such as structure-amplitude-location (SAL) are widely used, especially in validation of operational forecasts (e.g. Hersbach, 2000; Wernli et al., 2008).

1.2 The problem with mean-square error

Such metrics are very useful for model intercomparison and process-level diagnostics; however, they do not address the central issue in training data-driven models – that one must choose a differentiable loss function that somehow encodes the assumed data-generating mechanism. In practice, most deterministic regression models for precipitation are trained with root-mean-square error or mean-square error (RMSE/MSE; hereafter we will simply refer to both as MSE). This implicitly assumes Gaussian residuals with symmetric errors, which is violated by precipitation. As a result, models trained on MSE under-penalise failures on rare heavy events, allow predictions of negative rainfall (which must be corrected post hoc), and mishandle the mass at zero.

Despite these obvious shortcomings, MSE remains in near-ubiquitous use for data-driven precipitation models. For example, in nowcasting (e.g., Lebedev et al., 2019; Xiang et al., 2020; Cambier van Nooten et al., 2023; Han et al., 2023; Kim et al., 2024; Ayzel and Heistermann, 2025; Cao et al., 2025). In statistical and data-driven downscaling, MSE is similarly standard as the pixel-wise training objective (e.g., Vandale et al., 2017; Kumar et al., 2021; Wang et al., 2021; Reddy et al., 2023; Chiang et al., 2024; Liu et al., 2025). Other precipitation-related learning tasks (e.g., merging datasets, data-driven discovery, and AI-based global weather forecast models) also default to MSE (e.g., Moraux et al., 2021; Hunt and Turner, 2024; Lam et al., 2023; Kochkov et al., 2024). A few studies have begun to explore alternatives, e.g. frequency-domain or correlation-aware losses for nowcasting (Yan et al., 2024), cross-entropy when modelling occurrences (e.g., Shi et al., 2015; Agrawal et al., 2019), or adversarial/transport-style objectives in downscaling (e.g., Kumar et al., 2023). However, these remain the exception rather than the rule.

1.3 Tweedie distributions

Against this backdrop, the Tweedie family offers an attractive compromise. As an exponential-dispersion family with variance function $V(\mu) = \mu^p$, Tweedie models provide a continuum connecting Gaussian ($p \rightarrow 0$), Poisson ($p \rightarrow 1$), and Gamma ($p \rightarrow 2$) limits, and for $1 < p < 2$ represent a compound Poisson–Gamma law that accommodates zero inflation together with continuous, positively supported amounts. This makes the Tweedie particularly appealing for precipitation, where we want to simultaneously model a frequent no-rain state alongside a skewed distribution of strictly positive intensities. Earlier methodological work has made likelihood-based fitting practical (e.g., Dunn, 2004), and applications to monthly and seasonal rainfall in different climates have demonstrated competitive performance and interpretability (e.g., Hasan and Dunn, 2010, 2011, 2012; Svensson et al., 2017; Suhaila, 2023; Dzupire et al., 2018). Nevertheless, Tweedie models remain underused relative to their relevance.

In this paper, we will argue that the training objective must be statistically consonant with the target variable. Because precipitation is non-negative, highly skewed, and frequently zero, a Gaussian-implied MSE is a poor default. In contrast, the Tweedie deviance (derived in Section 2) provides a coherent likelihood-based loss that accounts for these characteristics while remaining tractable for gradient-based training. Crucially, the Tweedie index parameter p can be estimated empirically (e.g., from the variance-mean power law) and adapted to different temporal aggregations, yielding a family of training targets that correctly interpolate between Poisson-like and Gamma-like behaviours rather than incorrectly imposing an a priori Gaussian assumption. The paper is laid out as follows. In Sec. 2, we explain in more detail the features of the Tweedie that make it desirable. We then demonstrate that Gaussianity is not a reasonable baseline for precipitation distributions across scales and show how to estimate the Tweedie power parameter from data (Sec. 4.1), before quantifying the impact of replacing MSE with Tweedie deviance in two common learning tasks for precipitation: downscaling (Sec. 4.2) and nowcasting (Sec. 4.3).

2 Important features of the Tweedie distribution

An introduction to the Tweedie distribution and its advantages follows. Throughout, we will use y to mean the observation, θ to mean the canonical parameter of the distribution, κ to mean the cumulant function, and $\phi > 0$ to mean the dispersion. For any power p (with $p \neq 1, 2$) the Tweedie belongs to the exponential dispersion family:

$$f_Y(y; \theta, \phi) = a(y, \phi, p) \exp \left[\frac{y\theta - \kappa(\theta)}{\phi} \right], \quad (1)$$

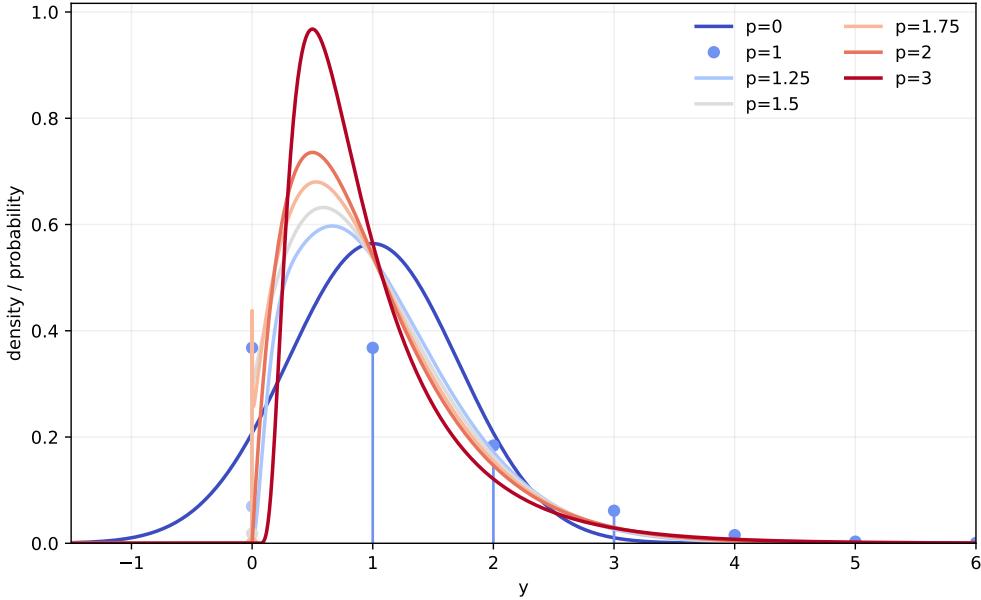


Figure 1: Tweedie family PDFs for fixed mean and dispersion. Curves show $f_Y(y)$ for $\mu = 1$ and $\phi = 0.5$ across powers $p \in \{0, 1, 1.25, 1.5, 1.75, 2, 3, 4\}$. Special cases: $p = 0$ is $\mathcal{N}(\mu, \phi)$ (solid line), $p = 1$ is Poisson(μ) (probability mass function shown as stems), $p = 2$ is Gamma, $p = 3$ is inverse-Gaussian. For $1 < p < 2$ the distribution is compound Poisson–Gamma with a point mass at zero (stem height $\exp\{-\mu^{2-p}/[\phi(2-p)]\}$) and a continuous density on $y > 0$. For $p > 1$ support is $y > 0$; the variance scales as $\text{Var}(Y) = \phi \mu^p$. Continuous lines are densities; stems indicate discrete probability mass.

whose variance function is

$$\text{Var}(Y) = \phi V(\mu), \quad V(\mu) = \mu^p. \quad (2)$$

The mean μ and canonical parameter θ are then related through:

$$\theta = \frac{\mu^{1-p}}{1-p}, \quad \kappa(\theta) = \frac{\mu^{2-p}}{2-p}, \quad (3)$$

and the normalising factor, $a(\cdot)$, is whatever constant in y is required to make the density function integrate to one.

What we are really interested in, however, is the deviance. For a single observation, y , the unit deviance is the scaled log-likelihood ration between the saturated (i.e., perfectly fit) model and model with mean μ :

$$d(y, \mu) = 2 [\ell(y; \theta_y, \phi) - \ell(y; \theta_\mu, \phi)], \quad (4)$$

where

$$\ell(y; \theta, \phi) = \frac{y\theta - \kappa(\theta)}{\phi} + \log a(y, \phi, p), \quad (5)$$

and θ_y is the canonical parameter corresponding to mean y (similarly for μ). Because the normalising term $\log a(y, \phi, p)$ does not depend on θ , it cancels in the difference, and the unit deviance (i.e. without the $1/\phi$ factor) is

$$d(y, \mu) = 2 [y(\theta_y - \theta_\mu) - (\kappa(\theta_y) - \kappa(\theta_\mu))]. \quad (6)$$

The difference terms can be computed through integration, i.e.:

$$\theta_y - \theta_\mu = \int_\mu^y \frac{d\theta}{dt} dt = \int_\mu^y \frac{dt}{V(t)} \quad (7)$$

and

$$\kappa(\theta_y) - \kappa(\theta_\mu) = \int_{\theta_\mu}^{\theta_y} \kappa'(\theta) d\theta = \int_\mu^y \frac{t}{V(t)} dt, \quad (8)$$

where we have used the identity $\mu = \mathbb{E}(y) = \kappa'(\theta)$ which holds for all members of the exponential distribution family. Substituting these values into Eq. 6 then gives a single integral form for the unit deviance of any exponential distribution function:

$$d(y, \mu) = 2 \int_\mu^y \frac{y-t}{V(t)} dt. \quad (9)$$

Thus, for the Tweedie, we have:

$$\begin{aligned} d_p(y, \mu) &= 2 \int_\mu^y (y-t) t^{-p} dt \\ &= 2 \left[y \int_\mu^y t^{-p} dt - \int_\mu^y t^{1-p} dt \right] \\ &= 2 \left[\frac{y^{1-p} - \mu^{1-p}}{1-p} - \frac{y^{2-p} - \mu^{2-p}}{2-p} \right] \\ &= 2 \left[\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y \mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p} \right], \end{aligned} \quad (10)$$

for $p \neq 1, 2$. The total deviance for a sample, D , is then obtained in the usual way, $D = \frac{1}{\phi} \sum_i d_p(y_i, \mu_i)$.

2.1 Recovering the Gaussian ($p \rightarrow 0$)

Let $p = \varepsilon$ with $\varepsilon \rightarrow 0$. We then expand Eq. 3 as Taylor series:

$$\begin{aligned} \theta &= \frac{\mu^{1-\varepsilon}}{1-\varepsilon} = \mu \left[1 - \frac{\varepsilon}{2} (\ln \mu) \right] + O(\varepsilon^2), \\ \kappa(\theta) &= \frac{\mu^{2-\varepsilon}}{2-\varepsilon} = \frac{1}{2} \mu^2 + O(\varepsilon). \end{aligned} \quad (11)$$

Keeping only leading terms and substituting back into Eq. 1, we obtain the expected probability density function:

$$f_Y(y; \mu, \phi) \longrightarrow \underbrace{(2\pi\phi)^{-1/2}}_{a(y, \phi, 0)} \exp\left[-\frac{(y-\mu)^2}{2\phi}\right], \quad (12)$$

where a is determined through integration.

The Gaussian deviance, with $p \rightarrow 0$ and $V(\mu) = 1$, is then given by Eq. 9:

$$d_0(y, \mu) = 2 \int_\mu^y (y-t) dt = (y-\mu)^2, \quad (13)$$

which is, of course, the mean squared error.

2.2 Recovering the Poisson ($p \rightarrow 1$)

Let $p = 1 + \varepsilon$, again with $\varepsilon \rightarrow 0$. Then

$$\begin{aligned} \theta &= \frac{\mu^{-\varepsilon}}{-\varepsilon} = \ln \mu + O(\varepsilon) \\ \kappa(\theta) &= \frac{\mu^{1-\varepsilon}}{1-\varepsilon} = \mu + O(\varepsilon). \end{aligned} \quad (14)$$

Keeping only leading terms and substituting back into Eq. 1, we obtain the probability density function:

$$f_Y(y; \mu) = \underbrace{\frac{1}{y!}}_{a(y, \phi, 1)} \mu^y e^{-\mu}. \quad (15)$$

To obtain the deviance, we again use Eq. 9:

$$d_1(y, \mu) = 2 \int_\mu^y \frac{y-t}{t} dt = 2[y \ln(y/\mu) - (y-\mu)], \quad (16)$$

which is, of course, not the mean squared error.

2.3 Recovering the Gamma ($p \rightarrow 2$)

Let $p = 2 - \varepsilon$ with $\varepsilon \rightarrow 0$. From Eq. 1, we have:

$$\begin{aligned} \theta &= \frac{\mu^{-(1-\varepsilon)}}{-(1-\varepsilon)} = -\frac{1}{\mu} + O(\varepsilon) \\ \kappa(\theta) &= \frac{\mu^\varepsilon}{\varepsilon} = \ln \mu + O(\varepsilon). \end{aligned} \quad (17)$$

The normalising constant is rather more complicated than for $p = 0$ or $p = 1$:

$$a(y, \phi, 2) = \frac{y^{\frac{1}{\phi}-1}}{\Gamma(\frac{1}{\phi}) \phi^{\frac{1}{\phi}}}, \quad (18)$$

giving

$$f_Y(y; \mu, \phi) = \frac{1}{\Gamma(\frac{1}{\phi})} \left(\frac{y}{\phi \mu} \right)^{\frac{1}{\phi}-1} \frac{e^{-y/(\phi \mu)}}{\mu}. \quad (19)$$

The gamma distribution is more commonly written in terms of a shape parameter, $k = \frac{1}{\phi}$, and a scale parameter, $\beta = \frac{\mu}{k} = \phi \mu$. Thus we return the commonly known density function:

$$\text{Gamma}(y; k, \beta) = \frac{y^{k-1} e^{-y/\beta}}{\Gamma(k) \beta^k}. \quad (20)$$

Returning to Eq. 9, the deviance is then given by:

$$d_2(y, \mu) = 2 \int_\mu^y \frac{y-t}{t^2} dt = 2 \left[\frac{y-\mu}{\mu} - \ln\left(\frac{y}{\mu}\right) \right] \quad (21)$$

which is, of course, not the mean squared error.

3 Data and models

3.1 Sources of precipitation data

3.1.1 University of Reading observatory

We use rain-gauge observations from the University of Reading Atmospheric Observatory on the Whiteknights campus (United Kingdom). The site records precipitation with a tipping-bucket gauge connected to a METFiDAS-3 data logger, and provides WMO-standard processed output at 5-minute resolution from September 2014 onwards (details and additional instrumentation are described at <https://www.met.reading.ac.uk/observatorymain/precipitation.html>).

For this study, we work with the 5-minute series in units of mm per interval. Basic quality control is applied to ensure physical values: negative amounts are set to zero and non-finite entries are treated as missing. No wind-related undercatch correction or bias adjustment is applied. When accumulating to coarser temporal scales (30-min, hourly, daily, weekly, monthly), totals are formed by summation of the 5-min amounts within each interval. Intervals with missing sub-samples are excluded.

These gauge data provide a point-scale benchmark with high zero frequency at short accumulation times, which we will use for assessing the suitability of Tweedie-family losses and for contrasting behaviour across temporal aggregation (Section 4.1). Data can be freely downloaded from <https://metdata.reading.ac.uk/cgi-bin/MODE3.cgi>.

3.1.2 GPM-IMERG

GPM-IMERG (Global Precipitation Mission—Integrated Multi-satellite Retrievals for GPM; Huffman et al., 2020) has global coverage at a half-hourly, 0.1° resolution, starting June 2000 and continuing to the present day. Over the tropics, GPM-IMERG primarily ingests retrievals from (for 2000–2014) the defunct Tropical Rainfall Measuring Mission (Kummerow et al., 1998, 2000) 13.8 GHz precipitation radar and microwave imager (Kozu et al., 2001) and (for 2014–) the Global Precipitation Measurement (GPM; Hou et al., 2014) Ka/Ku-band dual-frequency precipitation radar. When an overpass is not available, precipitation is estimated by calibrating infrared measurements from geostationary satellites. While GPM IMERG performs well when compared against gauge-based products, performance falls at higher elevations or when quantifying extreme precipitation events (Prakash et al., 2018). Data were downloaded from <https://>

```
//disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_
06/summary?keywords=%22IMERG%20final%22.
```

3.1.3 ERA5

The ECMWF ERA5 reanalysis (Hersbach et al., 2020) provides global coverage at hourly resolution from 1940 to present on a 0.25° grid, with 37 pressure levels from 1000 to 1 hPa, plus single-level (surface, near-surface, or column-integrated) variables. It assimilates a wide range of conventional and satellite observations via 4D-Var. Many hydrometeorological quantities (including precipitation) are model diagnostics produced by the forecast model within the reanalysis rather than being directly assimilated.

For this study we use the single-level total precipitation field. Further preprocessing details (domain, splits, coarsening, thresholds) are given in App. A.3. Data were downloaded using the dedicated API <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels>.

3.2 Outline of model architectures

3.2.1 Downscaling

We train a conditional diffusion model to map coarse (1°) precipitation to the original 0.25° ERA5 precipitation. The score network is a U-Net with group normalisation and sinusoidal time embeddings. Model input is the high-resolution noise-added field concatenated with (i) the upsampled coarse precipitation and (ii) a static orography channel. We follow a linear noise schedule with $T=1000$ steps and sample with a deterministic denoising diffusion implicit model scheme (Song et al., 2020). Targets are non-negative, so we pass the network output through a softplus to enforce $\mu \geq 0$. We then compare loss functions on exactly the same architecture: (a) RMSE on the target x_0 (i.e., the clean 0.25° precipitation field) ; and (b) Tweedie deviance (Sec. 2) with power p estimated from the training set via a variance-mean power law (Sec. 4.1) on the same target. More granular details of the setup are given in App. A.3.

3.2.2 Nowcasting

For 30-min IMERG nowcasting over the Kolkata domain we use a residual ConvLSTM (Shi et al., 2017). This comprises two stacked convolutional LSTM blocks with a skip connection, batch normalisation, and a shallow (i.e., single layer) convolutional head. As in the downscaling model, we use softplus to ensure non-negativity. Inputs are the previous $L=4$ frames; the model predicts the next frame, and longer horizons are obtained by running the model autoregressively. As before, we train the same architecture with different losses: (a) RMSE and (b) Tweedie deviance. More granular details of the setup are given in App. A.2.

4 Case study applications

We will now present three case studies. The first will demonstrate that the Gaussian is not a reasonable *a priori* assumption for the distribution of precipitation. The second and third will demonstrate that nontrivial gains can be made in the skill of models used for common tasks (downscaling and nowcasting respectively) when the loss function is changed from RMSE to an appropriately constrained Tweedie deviance.

4.1 Goodness-of-fit to observations

Figure 2 summarises how the Tweedie power parameter, p , varies with accumulation period across our three independent datasets. A clear trend emerges, in that at short accumulations the process is close to the Poisson limit ($p \rightarrow 1$), while longer accumulations progressively drive p upwards towards the Gamma limit ($p \rightarrow 2$). In particular, the 5-min gauge data from Reading is very close to a Poisson distribution ($p = 1.18$), whereas the monthly totals in both ERA5 and IMERG have confidence intervals overlapping $p = 2$.

At the same time, there is a notable spread between datasets at the same accumulation period. For example, for daily totals, the estimated powers span roughly $p \in [1.57, 1.85]$. This inter-dataset variability is smaller than the systematic increase with accumulation period, but it is not negligible and cautions us against assuming a single p for a given timescale.

Including ERA5 for Kolkata shows the additional importance of location. Even with the same preprocessing, the daily estimates differ considerably between sites (Kolkata $p = 1.629$, Beijing $p = 1.852$). This partly reflects differences in storm organisation and frequency (e.g., convective storms embedded in the summer monsoon versus mixed regimes in the mid-latitudes), while the difference between IMERG and ERA5 at Kolkata indicates that dataset characteristics also contribute.

There are two practical conclusions from this. Firstly, none of the precipitation series shown here are remotely Gaussian at any accumulation: a Gaussian-consistent loss would require $p \rightarrow 0$, whereas all estimates sit in the Poisson-Gamma regime ($1 < p < 2$). Secondly, p is not knowable *a priori*. Instead, it must first be estimated from the training data at the target temporal scale and carried through model training and evaluation so that the deviance matches the problem.

Finally, we note that weekly accumulations consistently have lower p than daily accumulations. This holds for all datasets, but it is unclear why. Further sensitivity tests are required.

4.2 Use in training a downscaling model

We now consider the relative performance of the downscaling model (Sec. 3.2.1) trained to minimise RMSE and Tweedie deviance respectively. Randomly-selected precipitating case studies from August in our testing period (Fig. 3) reveal mixed results. The first case (top row) has light precipitation in the northwest of the domain, which the Tweedie model overestimates and the

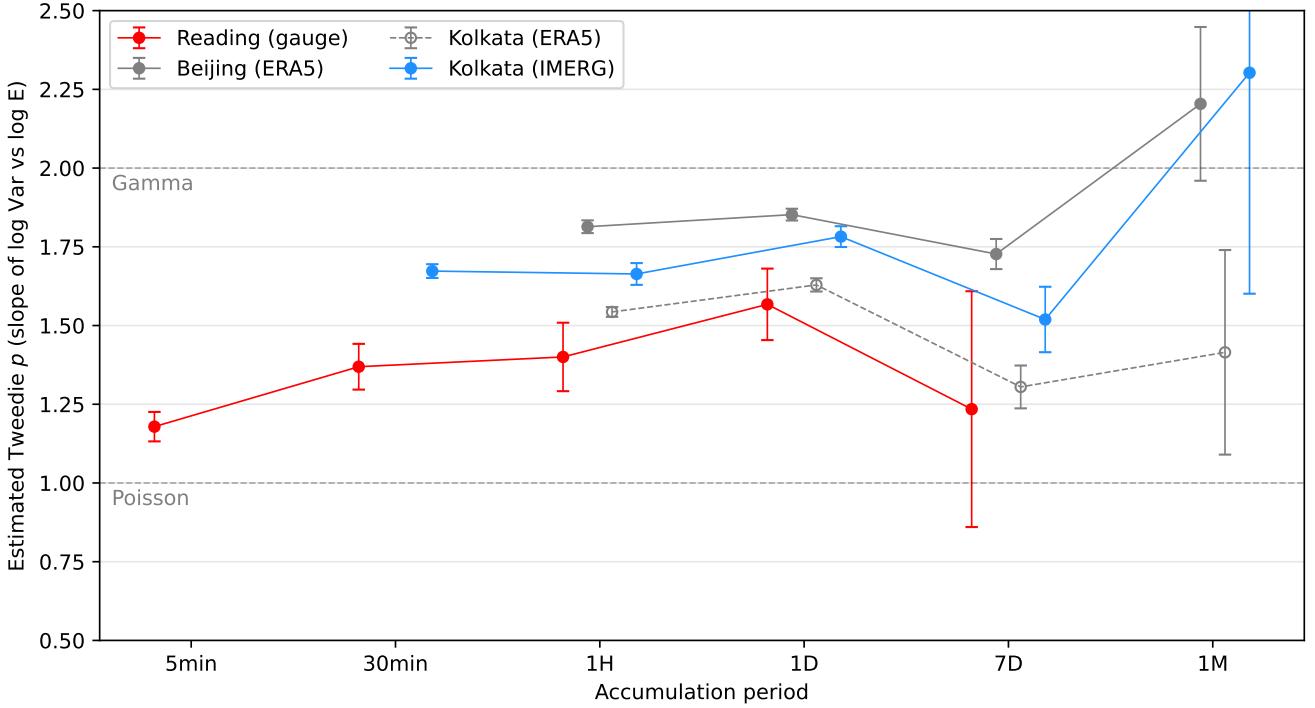


Figure 2: Estimated Tweedie power parameter p as a function of accumulation period for four datasets: Reading gauge (red), ERA5 at Beijing (grey), ERA5 at Kolkata (dashed grey) and IMERG at Kolkata (blue). For each dataset and accumulation period, totals are formed by summing raw values over the period and p is estimated from the power-law relation $\log \text{Var}(Y) = c + p \log E[Y]$ using ordinary least squares on non-overlapping blocks. Block lengths are: 10 days (5-min gauge), 15 days (30-min series), 30 days (hourly and daily), 16 weeks (weekly), and 12 months (monthly). Error bars show 95% Wald confidence intervals from the ordinary least squares slope standard error.

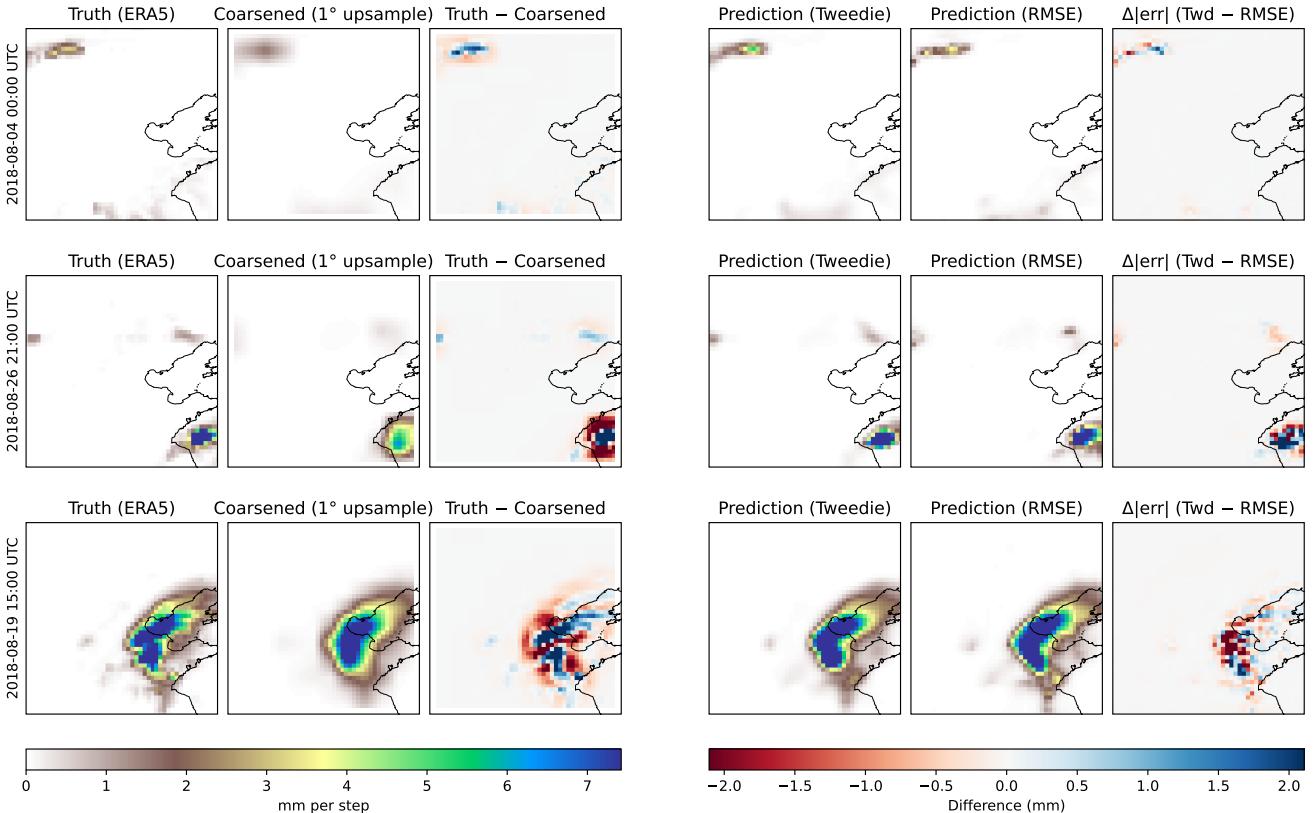


Figure 3: Performance of Tweedie vs RMSE diffusion downscaling for three randomly selected test cases. Column 1 shows the actual precipitation in ERA5; column 2 shows the ERA5 precipitation coarsened to 1° , i.e., the model input; and column 3 shows their difference. Column 4 and 5 show the predictions from the Tweedie and RMSE models respectively; and column 6 shows the difference in the mean absolute error between the models such that red (i.e., negative) values indicating a smaller error in the Tweedie model.

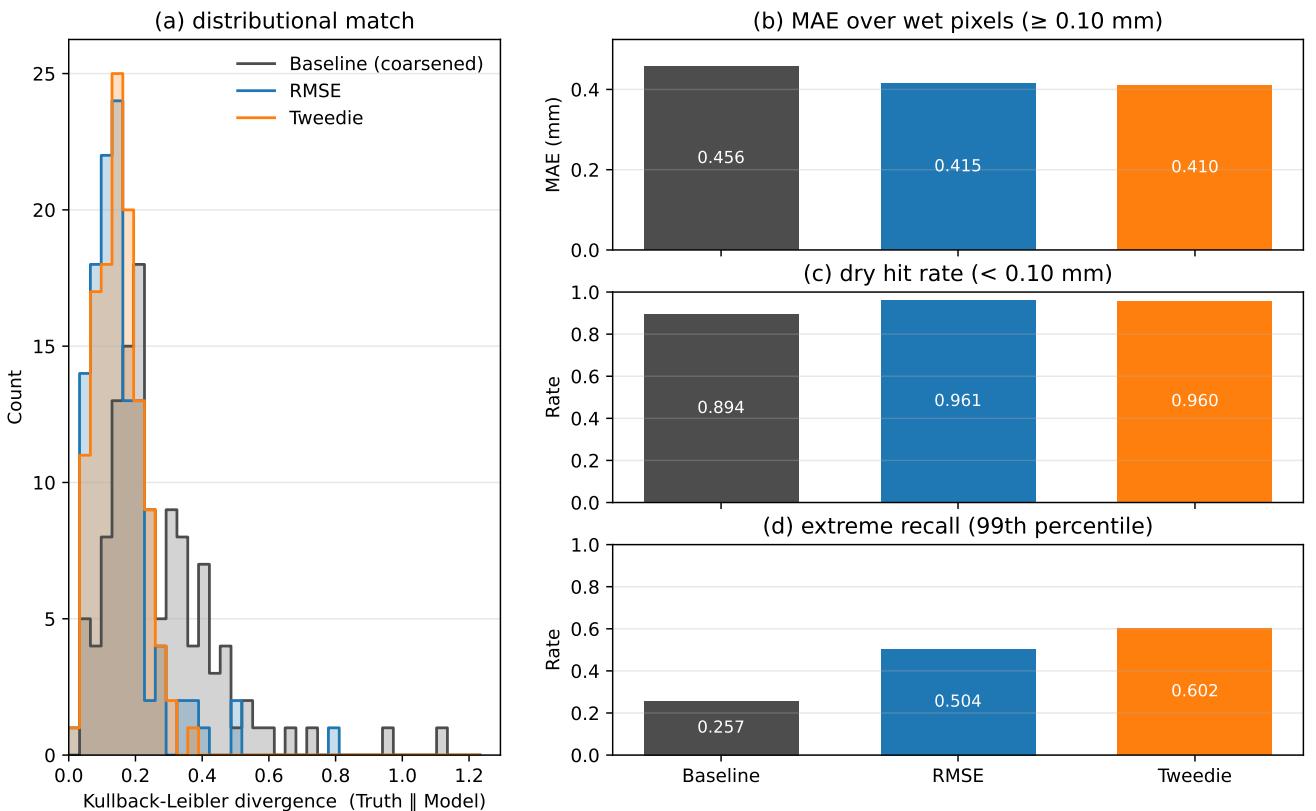


Figure 4: Comparison of Tweedie vs RMSE diffusion downscaling over the August 2018 Beijing test period. (a) Distribution of per-timestep Kullback–Leibler divergence $D_{\text{KL}}(\text{Truth} \parallel \text{Model})$ computed from histogrammed pixels (with ERA5 as truth). (b) mean absolute error over wet pixels (truth $\geq 0.1 \text{ mm hr}^{-1}$); (c) dry hit rate = $\Pr(\text{pred} < 0.1 \text{ mm hr}^{-1} \mid \text{truth} < 0.1 \text{ mm hr}^{-1})$; (d) extreme recall at the 99th-percentile (computed from truth over the period). Blue bars/curves show the RMSE-trained model; orange shows the Tweedie-trained model; grey shows the scores achieved by simply using the 1° coarsened data (i.e., no model) as a baseline. All metrics average over valid grid cells/times.

RMSE model underestimates. In the second case, a small but more intense storm in the southeast of the domain, both models correctly predict the high intensity and storm morphology, but the RMSE model has, in general, a lower mean absolute error. In the third case, a very intense and large storm, both models again predict the intensity and storm morphology correctly, but the Tweedie model has a lower overall absolute error.

Extending the analysis to the whole month of August 2018 (Fig. 4), we see that the two models perform similarly. We first compute, for each output, the Kullback-Leibler divergence between the output and the ‘true’ ERA5. The two distributions are virtually indistinguishable, and are not significantly different under a two-sample Kolmogorov-Smirnov test. This also holds for the MAE on wet pixels (RMSE: 0.415; Tweedie: 0.410) and the dry hit rate (RMSE: 0.961; Tweedie: 0.960). Again, the differences are not statistically significant when bootstrap tested. However, we do find a significant difference in 99th percentile recall, which is 0.504 for the RMSE model and 0.602 for the Tweedie model, a significant improvement.

So, although the two downscaling models have very similar performance across a range of metrics, we see a 19% improvement in predicting extreme rainfall with the Tweedie model over the RMSE model, due to better representation of the right tail in the loss function of the former.

4.3 Use in training a nowcasting model

Finally, we consider the respective performances of the Tweedie- and RMSE-trained nowcasting models, described in Sec. 3.2.2. Case study maps for a four-timestep rollout (Fig. 5) show that the Tweedie model is able to capture the structural features of the rainfall field, as well as broad characteristics of its evolution, including movement, growth and decay, across a range of scales. We are interested in the relative performance of the two models, shown in the rightmost two columns of Fig. 5. For both timesteps, and across all three cases, we see that the Tweedie model has lower error than the RMSE model at almost all locations. This includes both heavy rain and light rain events – although we note that for the heaviest rainfall case (bottom row), RMSE and Tweedie have a roughly balanced performance.

We evaluate the model performances over the same test month (August 2018; Fig. 6), chosen to cover some of the summer monsoon. We first compute, for each $t + 1$ output initialised from within the test period, the KL divergence between the output and the ‘true’ IMERG. Evaluated over all timesteps, this gives a distribution of KL divergences (Fig. 6a) which indicates a better performance from the Tweedie model. There are two desirable properties here. Firstly, that the median of the Tweedie KL distribution is closer to zero – indicating that more timesteps have a distribution of predicted rainfall closer to the observation than in the RMSE model. Secondly, that the right tail is considerably weaker in the Tweedie model – indicating fewer

busts. We next compute a set of forecast metrics for each of $t + 1$ through $t + 4$. As the models are only trained to produce predictions for $t + 1$, we use autoregression for subsequent timesteps, such that the prediction for $t + 4$ comprises three prediction timesteps and one observation. The mean absolute error over wet pixels (Fig. 6b) is consistently smaller in the Tweedie model than the RMSE model. The improvement is initially modest – from 0.452 mm at $t + 1$ to 0.444 mm (a reduction of 2%). However, by $t + 4$, the reduction is considerable at 16%. Similarly, the hit rate for dry pixels improves by about 2% (0.915 vs 0.896) in the Tweedie model for $t + 1$, but that improvement increases to about 13% (0.791 vs 0.701) at $t + 4$.

5 Conclusions

We have argued that root-mean-square error (RMSE) is a poor default objective for precipitation because it imposes Gaussian residuals on a target that is not remotely Gaussian, being semi-continuous, zero-inflated, non-negative, and heavy-tailed. As a statistically supported alternative, we advocate the Tweedie deviance, a likelihood-based, differentiable loss whose variance function $V(\mu) = \mu^p$ admits a compound Poisson–Gamma distribution for $1 < p < 2$. It thus naturally accommodates the point mass at zero and the skewed positive support observed for rainfall.

Empirically, we demonstrated that precipitation across three independent datasets is far from Gaussian at all practical temporal aggregations, with Tweedie power p lying firmly between the Poisson and Gamma limits, increasing with accumulation period towards the latter. Estimating p via the variance–mean power law yields values that are sensitive to both timescale and location. Using the same architectures and training setup, replacing RMSE with Tweedie deviance, we:

- produced comparable overall scores on downscaling but a materially better extreme recall (e.g., ~ 0.60 vs ~ 0.50 at the 99th percentile in the downscaling model), and
- improved nowcasting in a way that compounds underautoregressive rollout: small improvements at $t+1$ in wet-pixel MAE and dry hit rate increased to considerably by $t+4$, alongside a thinner right tail in the per-timestep KL divergence distribution (fewer busts).

This matters because the training objective shapes what the network learns. A Gaussian-implied loss underpenalises heavy events, allows negative precipitation, and ignores the mass at zero. The Tweedie deviance aligns the objective with observed precipitation statistics, while remaining simple to implement. Because the deviance is continuous in p , it provides a single framework that smoothly bridges short-interval, arrival-dominated behaviour and longer-interval accumulation statistics.

Our results use a single global p per experiment. Future work should quantify sensitivity to incorrect p , con-

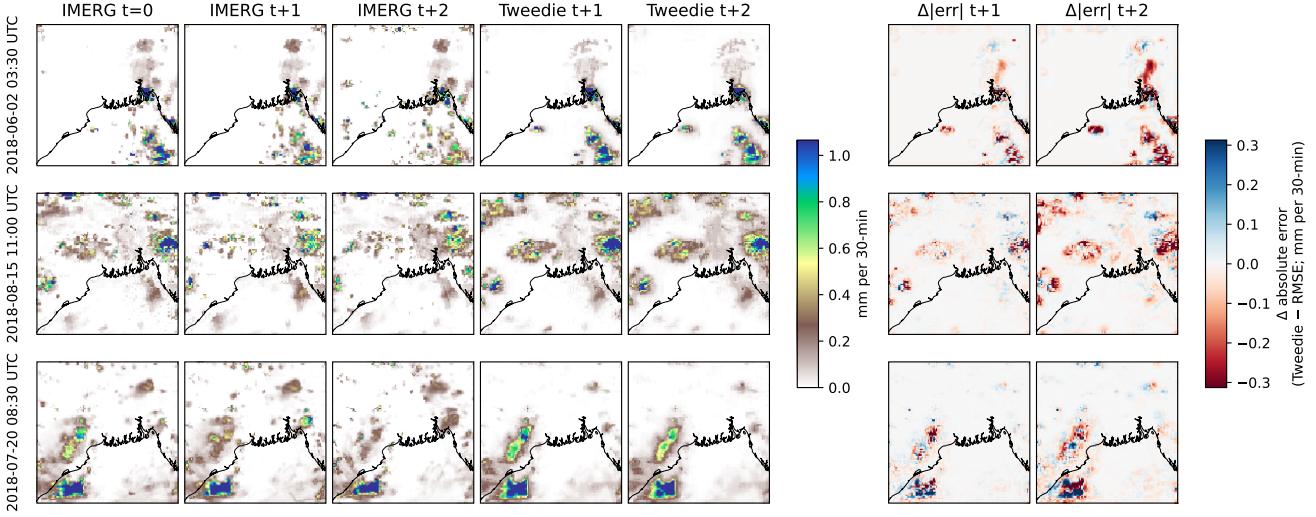


Figure 5: Performance of Tweedie vs RMSE ConvLSTM Kolkata nowcasts for three randomly selected test initialisations. Column 1 shows the last observed IMERG frame ($t=0$) that the models see; columns 2–3 show the subsequent half-hourly observations ($t+1, t+2$). Columns 4–5 show Tweedie nowcasts at $t+1$ and $t+2$ (with $t+2$ obtained autoregressively by feeding the $t+1$ prediction back as input). Columns 6–7 show the difference in mean absolute error such that red (i.e., negative) values indicating a smaller error in the Tweedie model.

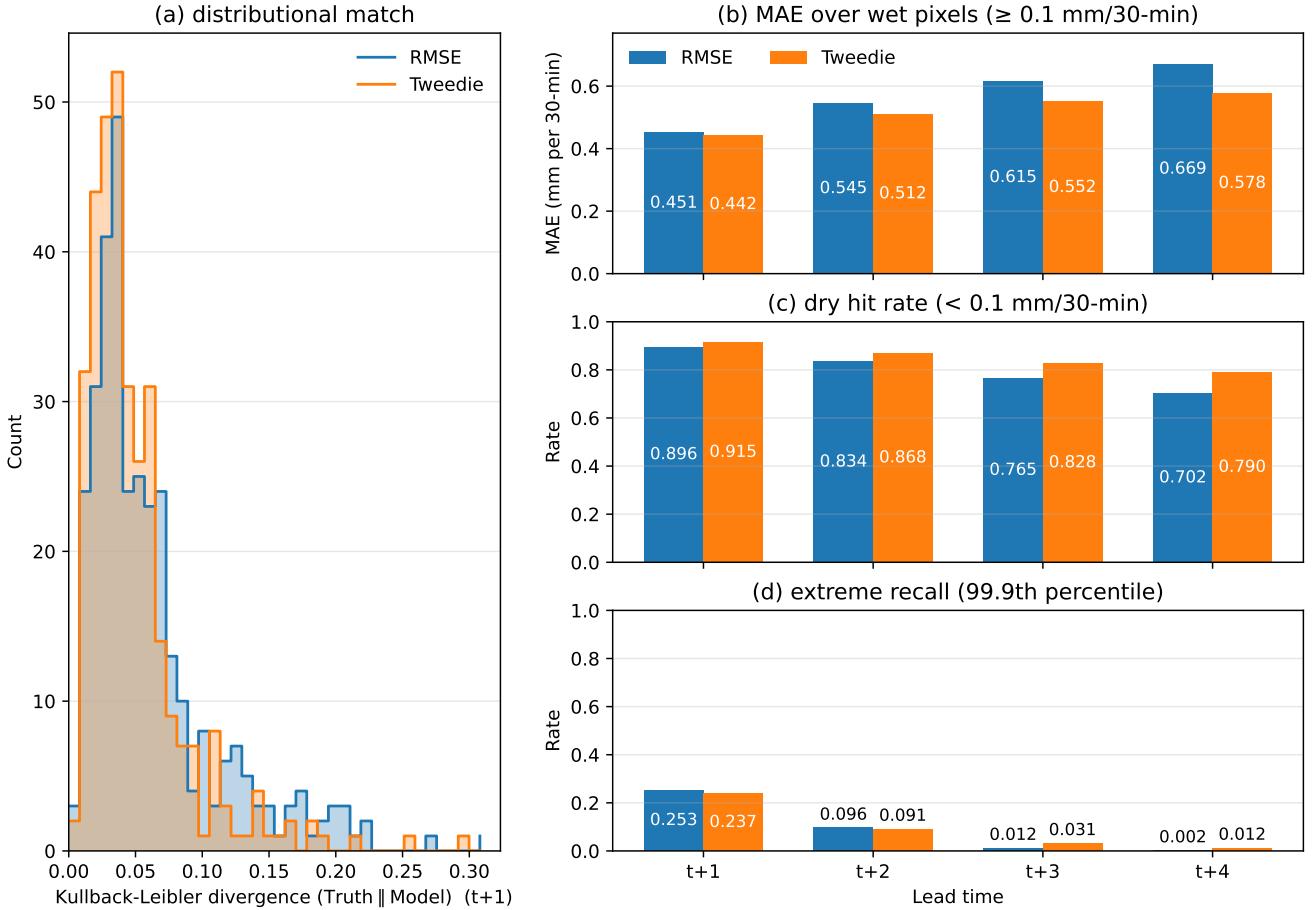


Figure 6: Comparison of Tweedie vs RMSE ConvLSTM nowcasts over the August 2018 Kolkata test period. (a) Distribution of $t+1$ Kullback–Leibler divergence $D_{KL}(\text{Truth} \parallel \text{Model})$ with IMERG as truth. (b–d) Mean scores vs lead time ($t+1$ to $t+4$) from autoregressive rollouts (each step feeds the previous prediction back as input): (b) mean absolute error over wet pixels (truth $\geq 0.1 \text{ mm per 30 min}$); (c) dry hit rate = $\text{Pr}(\text{pred} < 0.1 \text{ mm} \mid \text{truth} < 0.1 \text{ mm})$; (d) extreme recall at the 99.9th-percentile threshold (computed from truth over the period). Blue bars/curves show the RMSE-trained model; orange shows the Tweedie-trained model. All metrics average over valid grid cells/times.

sider spatially or seasonally varying p , and compare alternative estimators (e.g., profile-likelihood) against the blockwise variance–mean fit we used here. We would also like to explore extending these results to probabilistic training targets, which would allow quantification of model prediction uncertainty. Finally, the observed “weekly dip” in the p -aggregation curve needs a targeted analysis to distinguish finite-sample artefacts from genuine multiscale structure.

For deterministic precipitation learning tasks, we therefore recommend replacing RMSE with Tweedie deviance. p can be estimated from the training data at the target accumulation period, and non-negativity can be enforced at the output (although it may naturally arise from using the Tweedie deviance). This approach delivers equal-or-better performance overall and systematically improves behaviour where it matters most, extremes.

A Model definitions

A.1 Common conventions

All targets are non-negative precipitation in millimetres per model timestep (ERA5 hourly; IMERG half-hourly). To stabilise optimisation and preserve physical support we:

- enforce non-negativity at the output via a smooth positive mapping (using softplus);
- scale inputs and targets by a single scalar s , defined as the 99th percentile of the *training* target pixels over space–time;
- use AdamW (learning rate 2×10^{-4} , weight decay 10^{-4}), global gradient-norm clip at 1.0, early stopping on validation loss with a patience of 4 epochs, and an exponential moving average (EMA) of weights with decay 0.999 for validation/sampling;
- for Tweedie deviance, set dispersion $\phi = 1$ and estimate the power p once on the *training* split (for more details, see App. A.4).

A.2 Nowcasting model

We use a native 100×100 grid at 0.1° resolution centred on Kolkata ($22.57^\circ\text{N}, 88.36^\circ\text{E}$). Our training period is 2014-06-01 to 2017-05-31; validation is 2017-06-01 to 2017-12-31; and test is 2018-06-01 to 2018-08-31.

For the LSTM, we use a context length of $L = 4$ (i.e., the last 2 h). The target is the next frame at $t+1$. Stride-1 sliding windows form examples with input tensor shape $(L, H, W, 1)$ and target $(H, W, 1)$ in scaled units. Longer horizons $t+k$ are obtained autoregressively by feeding the $(t+k-1)$ prediction back as the next context frame.

The architecture is as follows. All convolutions use 3×3 kernels with “same” padding. The hidden width is 48 channels throughout. Block 1 is a ConvLSTM2D layer (filters 48, return full sequence) followed by batch normalisation and SpatialDropout3D with rate 0.10.

Block 2 is a second ConvLSTM2D layer (filters 48, return full sequence) + batch normalisation + SpatialDropout3D (also 0.10). A residual add combines the outputs of Blocks 1 and 2 (elementwise over the sequence). Finally we apply a third ConvLSTM2D (filters 48, return last frame) + batch normalisation; then a 3×3 convolution with 48 filters and SiLU activation, followed by a 1×1 convolution to 1 channel and a final softplus to ensure $\mu \geq 0$. Kernels use He-normal initialisation; biases are zero. Batch-norm scales/offsets start at 1/0 as standard.

For training, we use a batch size 16 and a maximum of 50 epochs with early stopping.

A.3 Downscaling model

ERA5 hourly precipitation is used, with negatives set to zero and non-finite values dropped. Target (“high-resolution”) is a native 50×50 grid at $\sim 0.25^\circ$ centred on Beijing ($39.90^\circ\text{N}, 116.41^\circ\text{E}$). A 1.0° conditioner is constructed by 4×4 block-mean coarsening the native grid followed by bilinear interpolation back to the native grid size. A static orography (geopotential) field, standardised to zero mean and unit variance is added as a second conditioning channel. Our training period is 1980-01-01 to 2016-12-31; validation is 2017-01-01 to 2017-12-31; and testing is over 2018-01-01 to 2018-12-31.

We train a conditional diffusion model that predicts the clean high-resolution field x_0 (in scaled units) from a noisy version x_t together with a conditioner (coarse-upscaled precipitation and orography). During training a timestep $t \in \{0, \dots, T-1\}$ is sampled, with $T = 1000$, linear $\beta_t \in [10^{-4}, 2 \times 10^{-2}]$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s \leq t} \alpha_s$; the forward process draws $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$.

The score network is a depth-2 encoder–decoder with GroupNorm and SiLU activations and a base width of 48 channels. A scalar continuous time $t_{\text{cont}} \in [0, 1]$ passes through a two-layer multilayer perceptron of width 4×48 and the projected embedding is added (broadcast) before each normalisation. The structure of the encoder is a DoubleConv($C_{\text{in}} \rightarrow b$) then $2 \times$ max-pool; DoubleConv($b \rightarrow 2b$) then $2 \times$ pool; bottleneck DoubleConv($2b \rightarrow 4b$). The structure of the decode is upsample $\times 2$; then concatenate skip from $2b$; DoubleConv($4b+2b \rightarrow 2b$); upsample $\times 2$; concatenate skip from b ; DoubleConv($2b+b \rightarrow b$); 1×1 convolution to 1 channel and final softplus to obtain $\tilde{x}_0 \geq 0$. Each DoubleConv is Conv $3 \times 3 \rightarrow$ GroupNorm(G) \rightarrow SiLU \rightarrow Conv $3 \times 3 \rightarrow$ GroupNorm(G) \rightarrow SiLU with time-embedding add-ins. G is the largest divisor of the channel count not exceeding 8 (to ensure exact divisibility). As before, kernels use He-normal initialisation; biases are zero; GroupNorm scales/offsets start at 1/0.

The channel stack used for training comprises $[x_t, \text{coarse-upscaled, orography}]$ (first two in scaled units; orography standardised). The target is the clean high-resolution field x_0 in scaled units.

We compare (i) root-MSE(\tilde{x}_0, x_0) and (ii) Tweedie

deviance $d_p(x_0, \tilde{x}_0)$ with p from App. A.4, both in scaled units and with a softplus applied. The batch size is 32 and we train up to 10 epochs with early stopping, maintaining an EMA (decay 0.999).

The generating model then has $K = 150$ evenly spaced integer timesteps from $T-1$ to 0 and $\eta = 0$. At each chosen t , we compute \tilde{x}_0 and $\hat{\epsilon} = (x_t - \sqrt{\bar{\alpha}_t} \tilde{x}_0) / \sqrt{1 - \bar{\alpha}_t}$, then jump to the previous chosen step t' via $x_{t'} = \sqrt{\bar{\alpha}_{t'}} \tilde{x}_0 + \sqrt{1 - \bar{\alpha}_{t'}} \hat{\epsilon}$. At $t' = 0$ return \tilde{x}_0 and rescale by s back to mm hr^{-1} .

A.4 Estimating the Tweedie power p

We estimate p using the variance–mean power law. For each pixel we create non-overlapping temporal blocks and fit $\log \text{Var}(Y) = c + p \log \mathbb{E}[Y]$ using ordinary least squares across blocks. Blocks are 30 days long (giving for IMERG 48×30 half-hourly samples and for ERA5 24×30 hourly samples). The estimate is then averaged across pixels and this value is fixed for validation and testing. The Tweedie dispersion parameter ϕ only scales the deviance and is set to 1 during training.

A.5 Evaluation metrics

At each time, let Z be the set of pixels and \hat{y} the prediction. Wet-pixel MAE is defined as the average $|\hat{y}_i - y_i|$ over $i \in W = \{i \in Z : y_i \geq \tau\}$. Dry hit rate is defined as $\Pr(\hat{y} < \tau | y < \tau)$. Extreme recall is defined as $\Pr(\hat{y} \geq u_q | y \geq u_q)$, where u_q is the q th percentile of truth over the evaluation period (nowcasting $q=99.9$, downscaling $q=99$).

Acknowledgments

KMRH is supported by a NERC Independent Research Fellowship (MITRE; NE/W007924/1).

References

- AghaKouchak, A., and N. Nasrollahi, 2010: Semi-parametric and parametric inference of extreme value models for rainfall data. *Water resources management*, **24** (6), 1229–1249.
- Agrawal, S., L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey, 2019: Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*.
- Amburn, S. A., A. S. I. D. Lang, and M. A. Buonaiuto, 2015: Precipitation forecasting with gamma distribution models for gridded precipitation events in eastern Oklahoma and northwestern Arkansas. *Weather and Forecasting*, **30** (2), 349–367.
- Ayzel, G., and M. Heistermann, 2025: Brief communication: Training of AI-based nowcasting models for rainfall early warning should take into account user requirements. *Natural Hazards and Earth System Sciences*, **25** (1), 41–47.
- Brier, G. W., and Coauthors, 1950: Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78** (1), 1–3.
- Cambier van Nooten, C., K. Schreurs, J. S. Wijnands, H. Leijnse, M. Schmeits, K. Whan, and Y. Shapovalova, 2023: Improving precipitation nowcasting for high-intensity events using deep generative models with balanced loss and temperature data: A case study in the Netherlands. *Artificial Intelligence for the Earth Systems*, **2** (4), e230017.
- Cao, Y., L. Chen, J. Wu, and J. Feng, 2025: Enhancing nowcasting with multi-resolution inputs using deep learning: Exploring model decision mechanisms. *Geophysical Research Letters*, **52** (4), e2024GL113699.
- Chiang, C.-H., and Coauthors, 2024: Climate downscaling: a deep-learning based super-resolution model of precipitation data with attention block and skip connections. *arXiv preprint arXiv:2403.17847*.
- Cho, H.-K., K. P. Bowman, and G. R. North, 2004: A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied meteorology*, **43** (11), 1586–1597.
- Cowpertwait, P. S., 1994: A generalized point process model for rainfall. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, **447** (1929), 23–37.
- Cowpertwait, P. S. P., C. G. Kilsby, and P. E. O’Connell, 2002: A space-time Neyman-Scott model of rainfall: Empirical analysis of extremes. *Water Resources Research*, **38** (8), 6–1.
- Dunn, P. K., 2004: Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, **24** (10), 1231–1239.
- Dzupire, N. C., P. Ngare, and L. Odongo, 2018: A Poisson-Gamma model for zero inflated rainfall data. *Journal of Probability and Statistics*, **2018** (1), 1012647.
- Ebert, E. E., L. J. Wilson, B. G. Brown, P. Nurmi, H. E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 forecast demonstration project. *Weather and forecasting*, **19** (1), 73–96.
- Gupta, V. K., and E. Waymire, 1998: Some mathematical aspects of rainfall, landforms, and floods. *Advanced series in statistical sciences and applied probability*, **7**, 129–172.
- Han, D., J. Im, Y. Shin, and J. Lee, 2023: Key factors for quantitative precipitation nowcasting using ground weather radar data based on deep learning. *Geoscientific Model Development*, **16**, 5895–5914, doi:10.5194/gmd-16-5895-2023, URL <https://gmd.copernicus.org/articles/16/5895/2023/>.

- Hasan, M. M., and P. K. Dunn, 2010: A simple Poisson–gamma model for modelling rainfall occurrence and amount simultaneously. *Agricultural and Forest Meteorology*, **150** (10), 1319–1330.
- Hasan, M. M., and P. K. Dunn, 2011: Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia. *International Journal of Climatology*, **31** (9), 1389–1397.
- Hasan, M. M., and P. K. Dunn, 2012: Understanding the effect of climatology on monthly rainfall amounts in australia using Tweedie GLMs. *International Journal of Climatology*, **32** (7).
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15** (5), 559–570.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146** (730), 1999–2049.
- Hogan, R. J., C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2010: Equitability revisited: Why the “equitable threat score” is not equitable. *Weather and Forecasting*, **25** (2), 710–726.
- Hou, A. Y., and Coauthors, 2014: The global precipitation measurement mission. *Bulletin of the American Meteorological Society*, **95** (5), 701–722.
- Huffman, G. J., and Coauthors, 2020: NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Algorithm Theoretical Basis Document (ATBD). Algorithm theoretical basis document, NASA Global Precipitation Measurement (GPM), Greenbelt, MD. URL https://gpm.nasa.gov/sites/default/files/2020-05/IMERG_ATBD_V06.3.pdf, version 06.3.
- Hunt, K. M. R., and A. G. Turner, 2024: Using interpretable gradient-boosted decision-tree ensembles to uncover novel dynamical relationships governing monsoon low-pressure systems. *Quarterly Journal of the Royal Meteorological Society*, **150** (758), 1–24, doi:<https://doi.org/10.1002/qj.4582>.
- Husak, G. J., J. Michaelsen, and C. Funk, 2007: Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **27** (7), 935–944.
- Katz, R. W., 1977: Precipitation as a chain-dependent process. *Journal of Applied Meteorology* (1962–1982), 671–676.
- Kim, W., C.-H. Jeong, and S. Kim, 2024: Improvements in deep learning-based precipitation nowcasting using major atmospheric factors with radar rain rate. *Computers & Geosciences*, **184**, 105 529.
- Kochkov, D., and Coauthors, 2024: Neural general circulation models for weather and climate. *Nature*, **632** (8027), 1060–1066.
- Kostinski, A. B., and A. R. Jameson, 1997: Fluctuation properties of precipitation. Part I: On deviations of single-size drop counts from the Poisson distribution. *Journal of the atmospheric sciences*, **54** (17), 2174–2186.
- Kozu, T., and Coauthors, 2001: Development of precipitation radar onboard the Tropical Rainfall Measuring Mission (TRMM) satellite. *IEEE transactions on geoscience and remote sensing*, **39**, 102–116, doi: 10.1109/36.898669, URL <http://dx.doi.org/10.1109/36.898669>.
- Kumar, B., K. Atey, B. B. Singh, R. Chattopadhyay, N. Acharya, M. Singh, R. S. Nanjundiah, and S. A. Rao, 2023: On the modern deep learning approaches for precipitation downscaling. *Earth Science Informatics*, **16** (2), 1459–1472.
- Kumar, B., R. Chattopadhyay, M. Singh, N. Chaudhari, K. Kodari, and A. Barve, 2021: Deep learning-based downscaling of summer monsoon rainfall data over Indian region. *Theoretical and Applied Climatology*, **143** (3), 1145–1156.
- Kummerow, C., W. Barnes, T. Kozu, J. Shiue, and J. Simpson, 1998: The Tropical Rainfall Measuring Mission (TRMM) sensor package. *Journal of Atmospheric and Oceanic Technology*, **15**, 809–817, doi:10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2, URL [http://dx.doi.org/10.1175/1520-0426\(1998\)015<0809:TTRMMT>2.0.CO;2](http://dx.doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2).
- Kummerow, C., and Coauthors, 2000: The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. *Journal of Applied Meteorology and Climatology*, **39** (12), 1965–1982, doi:10.1175/1520-0450(2001)040<1965:TSOTTR>2.0.CO;2, URL [http://dx.doi.org/10.1175/1520-0450\(2001\)040<1965:TSOTTR>2.0.CO;2](http://dx.doi.org/10.1175/1520-0450(2001)040<1965:TSOTTR>2.0.CO;2).
- Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382** (6677), 1416–1421.
- Lebedev, V., and Coauthors, 2019: Precipitation nowcasting with satellite imagery. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2680–2688.
- Liu, Y., Z. Li, G. Cao, Q. Wang, Y. Li, and Z. Lu, 2025: MSRGAN: A multi-scale residual GAN for high-resolution precipitation downscaling. *Remote Sensing*, **17** (13), 2281.
- Martinez-Villalobos, C., and J. D. Neelin, 2019: Why do precipitation intensities tend to follow gamma distributions? *Journal of the Atmospheric Sciences*, **76** (11), 3611–3631.
- Mittermaier, M. P., 2021: A “meta” analysis of the fractions skill score: The limiting case and implications for aggregation. *Monthly Weather Review*, **149** (10), 3491–3504.

- Moraux, A., S. Dewitte, B. Cornelis, and A. Munteanu, 2021: A deep learning multimodal method for precipitation estimation. *Remote Sensing*, **13** (16), 3278.
- Murata, A., S.-i. I. Watanabe, H. Sasaki, H. Kawase, and M. Nosaka, 2020: Assessing goodness of fit to a gamma distribution and estimating future projection on daily precipitation frequency using regional climate model simulations over Japan with and without the influence of tropical cyclones. *Journal of Hydrometeorology*, **21** (12), 2997–3010.
- Olivera, S., and C. Heard, 2019: Increases in the extreme rainfall events: Using the Weibull distribution. *Environmetrics*, **30** (4), e2532.
- Papalexiou, S. M., and D. Koutsoyiannis, 2013: Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, **49** (1), 187–201.
- Phoophiwa, T., P. Chomphuwise, T. Prahadchai, J.-S. Park, A. Apichottanakul, W. Theppang, and P. Busababodhin, 2024: Employing the generalized Pareto distribution to analyze extreme rainfall events on consecutive rainy days in Thailand's Chi watershed: implications for flood management. *Hydrology and Earth System Sciences*, **28** (4), 801–816.
- Prakash, S., A. K. Mitra, A. AghaKouchak, Z. Liu, H. Norouzi, and D. S. Pai, 2018: A preliminary assessment of GPM-based multi-satellite precipitation estimates over a monsoon dominated region. *Journal of Hydrology*, **556**, 865–876.
- Ramesh, N. I., C. Onof, and D. Xie, 2012: Doubly stochastic Poisson process models for precipitation at fine time-scales. *Advances in water resources*, **45**, 58–64.
- Reddy, P. J., R. Matear, J. Taylor, M. Thatcher, and M. Grose, 2023: A precipitation downscaling method using a super-resolution deconvolution neural network with step orography. *Environmental Data Science*, **2**, e17.
- Rho, H., and J. H. T. Kim, 2019: Modelling the entire range of daily precipitation using phase-type distributions. *Advances in Water Resources*, **123**, 210–224.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136** (1), 78–97.
- Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, 1987: Some models for rainfall based on stochastic point processes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, **410** (1839), 269–288.
- Segond, M.-L., C. Onof, and H. Wheater, 2006: Spatial-temporal disaggregation of daily rainfall from a generalized linear model. *Journal of Hydrology*, **331** (3-4), 674–689.
- Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, **28**.
- Shi, X., Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, 2017: Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, **30**.
- Song, J., C. Meng, and S. Ermon, 2020: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Strong, R., O. Borgstroem, R. Nathan, C. Wasko, and D. O'Shea, 2025: Global applicability of the Kappa distribution for rainfall frequency analysis. *Water Resources Research*, **61** (2), e2024WR039035.
- Suhaila, J., 2023: Tweedie models for Malaysia rainfall simulations with seasonal variabilities. *Journal of Water and Climate Change*, **14** (10), 3648–3670.
- Svensson, C., J. Hannaford, and I. Prosdocimi, 2017: Statistical distributions for monthly aggregations of precipitation and streamflow in drought indicator applications. *Water Resources Research*, **53** (2), 999–1018.
- Thom, H. C. S., 1958: A note on the gamma distribution. *Monthly weather review*, **86** (4), 117–122.
- Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Neiman, and A. R. Ganguly, 2017: DeepSD: Generating high resolution climate change projections through single image super-resolution. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1663–1672.
- Wang, F., D. Tian, L. Lowe, L. Kalin, and J. Lehrter, 2021: Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, **57** (4), e2020WR029308.
- Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, **136** (11), 4470–4487.
- Wilks, D. S., 1989: Rainfall intensity, the weibull distribution, and estimation of daily surface runoff. *Journal of Applied Meteorology and Climatology*, **28** (1), 52–58.
- Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of climate*, 1495–1501.
- Wilks, D. S., 1998: Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, **210** (1-4), 178–191.

Wilks, D. S., 1999: Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and forest meteorology*, **93** (3), 153–169.

Woolhiser, D. A., and J. Roldán, 1982: Stochastic daily precipitation models: 2. A comparison of distributions of amounts. *Water resources research*, **18** (5), 1461–1468.

Xiang, Y., J. Ma, and X. Wu, 2020: A precipitation nowcasting mechanism for real-world data based on machine learning. *Mathematical Problems in Engineering*, **2020** (1), 8408931.

Yan, C.-W., S. Q. Foo, V. H. Trinh, D.-Y. Yeung, K.-H. Wong, and W.-K. Wong, 2024: Fourier amplitude and correlation loss: Beyond using l2 loss for skillful precipitation nowcasting. *Advances in Neural Information Processing Systems*, **37**, 100 007–100 041.