Deep Extreme Transformer: Tackling Zero-Inflated Time Series for Precipitation Prediction

Wentao Gao[1], Xiongren Chen[1], Xiaojing Du[1], Wenjun Yu[2], Andres Mauricio Cifuentes Bernal[1], Ziqi Xu[3,*]

[1]Adelaide University, Adelaide, Australia
[2]Shanghai University of International Business and Economics, Shanghai, China
[3]RMIT University, Melbourne, Australia
a3166814@adelaide.edu.au, ziqi.xu@rmit.edu.au

Abstract

Rainfall forecasting presents a dual challenge: extreme *zero inflation*, where dry days dominate and obscure meaningful precipitation patterns, and pronounced *nonstationarity*, where climate dynamics evolve across time and regimes. We propose the **Deep Extreme Transformer (DET)**, a principled architecture that integrates statistical distribution modeling with neural sequence learning to address both issues simultaneously. DET augments the Transformer with a Tweedie distribution output head that unifies discrete zeros and continuous intensities, a fixed shared-weight mechanism that emphasizes rare but critical events in both attention and loss computation, and a Gaussian perturbation strategy that enhances learning stability without violating physical constraints. DET further incorporates nonstationary attention to adapt to evolving rainfall regimes. Extensive experiments on multi-decadal South Australian climate data demonstrate that DET consistently outperforms existing deep learning and statistical models across forecasting horizons. Our method provides an effective and generalizable framework for zero-inflated, shift-prone time series, bridging statistical rigor with deep temporal modeling in a unified and scalable design.

Introduction

Accurate rainfall prediction is crucial for agricultural planning, water resource management, and disaster preparedness in climate-vulnerable regions. However, a fundamental technical challenge undermines current forecasting systems: **zero inflation**, the overwhelming prevalence of zero rainfall observations that systematically biases prediction models away from detecting critical precipitation events.

The zero inflation problem in rainfall data is severe and pervasive (Schepen, Wang, and Robertson 2012). In drought-prone regions, dry days comprise 70-85% of all observations, creating datasets where the most frequent outcome provides virtually no predictive information about the rare but crucial events that communities need to anticipate. This extreme sparsity creates a fundamental modeling paradox: when trained on heavily zero-inflated data, machine learning approaches develop systematic biases toward predicting continued drought conditions (Wood et al. 2002; Wilby et al. 1998), failing to capture the onset of beneficial precipitation events that are essential for agricultural and water management decisions.

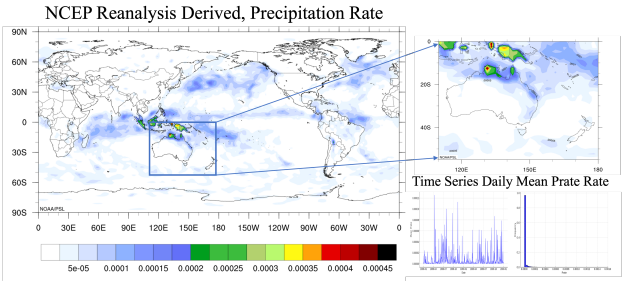

Figure 1: Rainfall time series demonstrating severe zero inflation and nonstationarity. The overwhelming presence of zero values masks critical precipitation events.

This zero inflation challenge is compounded by **nonstationarity**, where rainfall patterns exhibit continuously evolving statistical properties across seasons and climate regimes (Milly et al. 2008; Gao et al. 2024a). Models must simultaneously learn from extremely sparse, imbalanced data while adapting to shifting statistical foundations.

Current time series prediction models fail to adequately handle zero inflation (Coe and Stern 1982). Traditional statistical approaches like hurdle models (Mullahy 1986) and zero-inflated Poisson models (Lambert 1992) can manage sparse distributions but assume temporal independence (Cox 1955; Zucchini and MacDonald 2009), ignoring the sequential dependencies crucial for weather forecasting (Brockwell and Davis 2002).

More critically, modern deep learning approaches struggle catastrophically with zero inflation (Reichstein et al. 2019; Vandal et al. 2017). Transformer architectures (Vaswani et al. 2017), while excellent at capturing temporal dependencies, suffer from a fundamental flaw when faced with zero-inflated sequences: the self-attention mechanism becomes democratized across abundant zero observations (Clark et al. 2019; Rogers, Kovaleva, and Rumshisky 2020). Instead of learning to prioritize the few critical nonzero events, attention weights become nearly uniform, effectively diluting the model's ability to detect important precip-

*Corresponding author

itation signals. The Nonstationary Transformer (Liu et al. 2022), despite handling evolving statistical properties, remains fundamentally limited by this zero inflation problem.

This creates a critical gap in time series modeling: existing methods force a choice between handling data sparsity or capturing temporal dependencies (Han et al. 2019), but rainfall prediction requires both capabilities simultaneously. The zero inflation problem represents a fundamental barrier to developing reliable forecasting systems for sparse, sequential data (Wilks 2011).

To address this gap, we propose the **Deep Extreme Transformer (DET)**, specifically designed to handle severe zero inflation while preserving temporal modeling capabilities. Our main contributions are:

- **Theoretical analysis** proving why standard Transformers fail on zero-inflated sequences and establishing information-theoretic bounds for attention mechanisms.
- **Weighted attention mechanism** with fixed shared weights that restores information concentration by differentiating zero and non-zero observations in attention computation.
- **Tweedie distribution framework** providing unified modeling of both exact zeros and continuous positive precipitation values through principled statistical parameterization.
- **Gaussian perturbation strategy** for optimization stability that maintains statistical consistency while improving gradient properties in sparse regimes.

Related Work

Recent deep learning advances have shown significant promise for time series forecasting, with Transformer-based architectures particularly excelling due to their self-attention mechanisms and long-range dependency modeling capabilities (Zhou et al. 2021; Liu et al. 2022; Gao et al. 2024b; Nie et al. 2023; Kitaev, Kaiser, and Levskaya 2020; Yu et al. 2025; Xu et al. 2025). The Informer (Zhou et al. 2021) improves computational efficiency through sparse attention, while the Nonstationary Transformer (Liu et al. 2022) adapts to dynamic time series patterns. Earlier approaches using CNNs, RNNs, and GANs have also demonstrated success in capturing spatiotemporal dependencies (Norel et al. 2021), but these models uniformly treat zero and nonzero values, leading to suboptimal performance in detecting sparse rainfall events due to lack of zero inflation handling mechanisms.

Statistical approaches for zero-inflated data modeling have been extensively studied, with traditional methods like Zero-Inflated Poisson models (Lambert 1992) and specialized time series processes such as ZIPPAR (Maiti et al. 2014) addressing sparse count data through various estimation techniques. Recent work has explored zero-inflated Poisson regression for handling excess zeros and compared different zero-inflated models on modern datasets (Beveridge, Goldstein, and Chung 2024). The Tweedie distribution, with its unique Poisson-Gamma mixture property, has emerged as particularly suitable for zero-inflated continuous data in domains like insurance and precipitation modeling (Gao et al. 2025), though these statistical methods typically assume temporal independence and struggle with complex sequential patterns.

Hybrid approaches combining deep learning with statistical models represent an emerging trend for enhanced forecasting capabilities. (Shi, Dai, and Long 2021) demonstrated this fusion through a deep learning-based zero-inflated duration model for financial applications, while work in rainfall prediction combining statistical distributions with neural architectures remains limited. Our approach fills this gap by integrating Tweedie distribution parameterization with Transformer architectures, creating a unified framework that captures both temporal dependencies and zero inflation characteristics, addressing limitations of existing methods that force a choice between statistical rigor and sequential modeling capabilities.

Preliminaries

Problem statement

Our problem is essentially a time series forecasting task with a large number of zero values specifically in the target variable Y, which represents precipitation. We examine a time series of dimension m, denoted as $(\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathbf{R}^{T \times m}$. Our objective is to use historical data spanning h steps to predict the future k steps of the time series, expressed as $Y^* = g(\mathbf{X})$, where $\mathbf{X} \in \mathbf{R}^{h \times m}$ represents the input historical data and $Y^* \in \mathbf{R}^{k \times 1}$ denotes the predicted future precipitation values. In our work, we define $g(\cdot)$ as a transformer model.

Analysis of Zero-Inflation in Transformer

Zero-inflated data poses fundamental challenges for Transformer architectures. In standard self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

When target variables predominantly contain zeros, similarity scores $\frac{QK^T}{\sqrt{d_k}}$ between time steps become uniform due to minimal feature variation. This causes the softmax operation to produce nearly uniform attention weights, preventing the model from focusing on sparse but critical non-zero events.

The Nonstationary Transformer's de-stationary attention mechanism:

$$\text{Attn}(Q', K', V', \tau, \Delta) = \text{softmax}\left(\frac{\tau Q' K'^T + 1\Delta^T}{\sqrt{d_k}}\right)V' \tag{2}$$

suffers from the same limitation. Despite incorporating scaling factor τ and shifting vector Δ, the score matrix still exhibits low dynamic range when zeros dominate, leading to uniform attention distribution.

Figure 2 empirically demonstrates this phenomenon across varying zero ratios, showing how attention weights become increasingly uniform as zero-inflation increases.

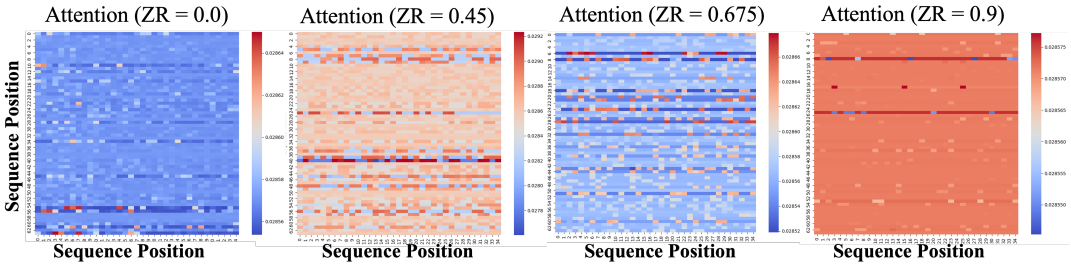

Figure 2: Comparison of attention heatmaps under varying zero ratios. The top row shows attention-weight heatmaps, where both axes denote sequence positions. Higher zero ratios lead to increasingly uniform and less concentrated attention patterns.

Proposed Method

Although the Nonstationary Transformer (Liu et al. 2022) addresses nonstationarity in time series data to some extent, it still faces fundamental challenges when dealing with zero-inflated data, particularly in capturing sparse non-zero events such as significant precipitation occurrences. To systematically overcome these limitations, we propose the DET, which integrates principled statistical modeling with neural sequence learning through four key innovations motivated by rigorous theoretical analysis. Figure 3 provides an overview of our complete architecture.

Problem Analysis: Why Standard Transformers Fail on Zero-Inflated Data

To understand why existing approaches struggle with zero-inflated sequences, we first establish the fundamental theoretical limitation that drives our design choices. In zero-inflated precipitation data, dry days comprise 70-85% of observations, creating a severe imbalance that systematically biases attention mechanisms.

Theorem 1 (Attention Weight Uniformity in Zero-Inflated Sequences)**.** *In zero-inflated sequences where $P(x_i = 0) = p_0$, as the zero-inflation ratio $p_0 \to 1$, the variance of attention similarity scores approaches zero, causing attention weights to converge to uniform distribution.*

Proof Sketch. Consider the similarity score matrix $S_{ij} = \frac{q_i k_j^T}{\sqrt{d_k}}$ in standard attention (Vaswani et al. 2017). When $p_0 \to 1$, most input vectors become sparse or zero-like. The variance of similarity scores becomes:

$$\text{Var}(S_{ij}) = \mathbb{E}[(q_i k_j^T)^2] - (\mathbb{E}[q_i k_j^T])^2 \to 0 \tag{3}$$

As variance diminishes, the softmax operation produces nearly uniform weights: $\lim_{p_0 \to 1} \text{softmax}(S_{ij}) = \frac{1}{T} + \epsilon$ where $\epsilon \to 0$. Full proof in Appendix A.1. □

This theorem reveals a critical flaw: as zero-inflation increases, Transformers lose their discriminative power. The attention mechanism becomes "democratized" across abundant zero observations, preventing the model from focusing on rare but critical precipitation events. This phenomenon is empirically demonstrated in Figure 2, where increasing zero ratios lead to increasingly uniform attention distributions.

Lemma 1 (Information Entropy Growth)**.** *The information entropy of standard attention increases with zero-inflation ratio:*

$$H_{standard} = -\sum_{j=1}^{T} \alpha_{ij} \log \alpha_{ij} \to \log T \text{ as } p_0 \to 1 \tag{4}$$

This means standard Transformers approach maximum uncertainty (minimum information) precisely when precision is most crucial for precipitation forecasting. The model carries maximum entropy when it should be most discriminative about rare events.

Solution 1: Weighted Attention for Information Concentration

Given the information loss identified in Theorem 1, we need a mechanism to restore the model's ability to focus on non-zero events. The core insight is that we must differentiate the importance of zero and non-zero observations in the attention computation itself. This motivates our weighted attention design.

Theorem 2 (Information Gain of Weighted Attention)**.** *Our weighted attention mechanism maintains information concentration by preserving the significance of non-zero events. The entropy of weighted attention satisfies $H_{weighted} < H_{standard}$, indicating more focused information distribution.*

Proof Sketch. By introducing differential weights $w_{ij} = \beta$ for non-zero events and $w_{ij} = \alpha$ for zero events (where $\beta > \alpha$), the weighted attention becomes:

$$\alpha_{ij}^w = \frac{\exp(S_{ij} \cdot w_{ij})}{\sum_k \exp(S_{ik} \cdot w_{ik})} \tag{5}$$

The exponential weighting amplifies non-zero positions while suppressing zero positions, creating a more concentrated distribution with lower entropy: $H_{\text{weighted}} = -\sum_j \alpha_{ij}^w \log \alpha_{ij}^w < H_{\text{standard}}$. Full proof in Appendix A.2 provided in the github. □

Based on this theoretical foundation, we implement a fixed shared weight mechanism. The weights are determined by past target values y_j in the decoder:

$$w_{i,j} = \begin{cases} \alpha, & \text{if } y_j = 0 \\ \beta, & \text{if } y_j > 0 \end{cases} \tag{6}$$

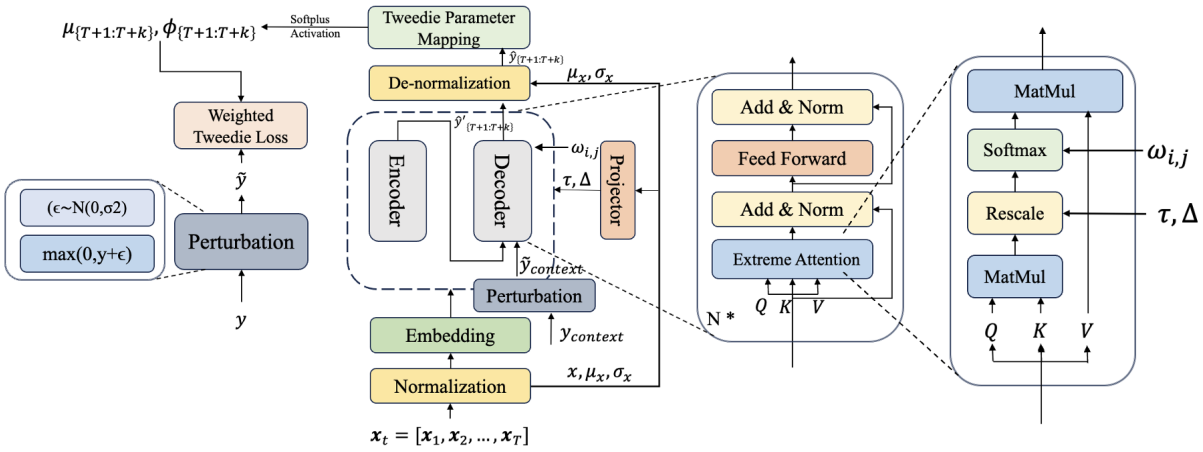

Figure 3: Architecture overview of the Deep Extreme Transformer. The model integrates four principled innovations to address zero-inflated precipitation forecasting: (1) Gaussian perturbation smooths zero values for stable optimization, (2) weighted attention with fixed shared weights (α, β) restores information concentration, (3) Tweedie parameter mapping provides unified zero-inflated modeling, and (4) weighted Tweedie loss ensures consistent optimization. The architecture seamlessly combines with nonstationary attention mechanisms (τ, Δ) to handle both zero-inflation and temporal nonstationarity simultaneously.

The weight ratio is carefully designed to balance contributions from zero and non-zero samples:

$$\alpha = \frac{p_{\text{non-zero}}}{p_{\text{zero}}} \cdot \beta \tag{7}$$

where p_{zero} and $p_{\text{non-zero}}$ are the empirical proportions of zero and non-zero observations in the dataset. For our South Australian precipitation data with 80% zeros, we have $p_{\text{zero}} = 0.8$, $p_{\text{non-zero}} = 0.2$, leading to $\alpha = 0.25\beta$ when $\beta = 1$.

The modified attention computation in the decoder becomes:

$$\text{Attention}(i,j) = \frac{\exp\left(\frac{\text{Score}(i,j)}{\sqrt{d_k}} \cdot w_{i,j}\right)}{\sum_k \exp\left(\frac{\text{Score}(i,k)}{\sqrt{d_k}} \cdot w_{i,k}\right)} \tag{8}$$

where $\text{Score}(i,j)$ is the standard dot-product similarity between query q_i and key k_j. This design ensures that attention weights for non-zero precipitation events receive higher emphasis, counteracting the uniformity bias identified in Theorem 1.

Theorem 3 (Attention Concentration Property)**.** *The weighted attention mechanism satisfies the concentration inequality:*

$$P\left(|\alpha_{ij}^w - \mathbb{E}[\alpha_{ij}^w]| \geq t\right) \leq 2\exp\left(-\frac{2t^2(\beta-\alpha)^2}{(\beta+\alpha)^2}\right) \tag{9}$$

This concentration bound guarantees that attention weights remain stable around their expected values, providing reliable focus on non-zero events with high probability.

Solution 2: Tweedie Distribution for Unified Zero-Inflated Modeling

While weighted attention addresses the discrimination problem, we also need an output distribution that naturally handles both exact zeros and continuous positive values. Standard regression assumes continuous targets with Gaussian noise, which fundamentally mismatches zero-inflated data characteristics. Traditional approaches force a choice between discrete (for zeros) and continuous (for positive values) modeling, but precipitation requires both simultaneously.

This motivates our Tweedie distribution parameterization, which provides a unified mathematical framework for zero-inflated continuous data. The Tweedie distribution is particularly suitable because it can generate exact zeros with positive probability while modeling continuous positive values.

Theorem 4 (Tweedie Distribution Optimality for Zero-Inflated Data)**.** *Among all distributions that can model both exact zeros and positive continuous values, the Tweedie distribution with shape parameter $1 < \rho < 2$ achieves maximum entropy under given mean and variance constraints.*

Proof Sketch. The Tweedie distribution belongs to the exponential family with form:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right) \tag{10}$$

where the natural parameter is $\theta = \frac{\mu^{1-\rho}}{1-\rho}$ and cumulant function is $b(\theta) = \frac{\mu^{2-\rho}}{2-\rho}$. When $1 < \rho < 2$, this represents a compound Poisson-Gamma mixture that uniquely combines discrete point mass at zero with continuous positive support while maximizing entropy. Full proof in Appendix A.3 provided in github. □

Tweedie Distributions for Zero-Inflated Precipitation Forecasting

The key properties that make Tweedie distributions ideal for our application are: (1) **Exact zero modeling**: The probabil-

ity of zero precipitation is explicitly modeled as

$$P(Y=0) = \exp\left(-\frac{\mu^{2-\rho}}{(2-\rho)\phi}\right) \tag{11}$$

(2) **Continuous positive values**: Non-zero values follow a continuous distribution with support $(0, \infty)$. (3) **Unified parameterization**: Both zero probability and positive value intensity are controlled by the same mean parameter μ. (4) **Exponential family properties**: Enable efficient parameter estimation and theoretical analysis.

We modify the Transformer's output layer to predict Tweedie distribution parameters rather than direct point estimates. The model's final hidden state $\hat{y} \in \mathbb{R}^{d_{\text{model}}}$ is mapped to distribution parameters through separate multi-layer perceptrons:

$$\mu = \text{softplus}(\text{MLP}_\mu(\hat{y})), \quad \phi = \text{softplus}(\text{MLP}_\phi(\hat{y})) \tag{12}$$

where MLP_μ and MLP_ϕ are two-layer feedforward networks with ReLU activations. The softplus activation ensures both parameters remain positive. The shape parameter ρ is treated as a learnable global parameter initialized at 1.5 and constrained to the interval $(1, 2)$ during training to maintain the Poisson-Gamma mixture property.

This parameterization enables the model to simultaneously learn: (1) when precipitation is likely to occur (μ controls zero probability), (2) how much precipitation is expected when it does occur (μ and ϕ control the positive value distribution), and (3) the variability in precipitation amounts (ϕ controls dispersion).

Solution 3: Gaussian Perturbation for Optimization Stability

Even with appropriate attention weighting and output distribution, training on zero-inflated data presents optimization challenges. The abundance of exact zeros creates non-differentiable point masses in the loss landscape, leading to unstable gradients and poor convergence. Traditional approaches either ignore this issue or use ad-hoc smoothing techniques without theoretical justification.

To address this systematically, we introduce a principled Gaussian perturbation strategy that maintains statistical consistency while improving optimization properties.

Lemma 2 (Statistical Consistency of Gaussian Perturbation)**.** *For zero-valued samples, applying truncated Gaussian perturbation $\tilde{y}_i = \max(0, y_i + \epsilon_i)$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ yields a perturbed distribution $q_\sigma(\tilde{y})$ that: (1) converges weakly to the original distribution as $\sigma \to 0$, (2) improves differentiability properties, and (3) preserves physical consistency through non-negativity constraints.*

Proof Sketch. Weak convergence $q_\sigma(\tilde{y}) \xrightarrow{w} p(y)$ as $\sigma \to 0$ follows from the continuous mapping theorem (Billingsley 2013). For differentiability, the truncated Gaussian smooths the non-differentiable point mass at zero, improving the Lipschitz constant of the gradient:

$$\|\nabla L_{\text{perturbed}}(\theta_1) - \nabla L_{\text{perturbed}}(\theta_2)\| \leq L_\sigma \|\theta_1 - \theta_2\| \tag{13}$$

where $L_\sigma < L_{\text{original}}$. The max operation preserves precipitation semantics by ensuring non-negativity. Full proof in Appendix A.4. □

The perturbation scale is chosen adaptively based on the data characteristics:

$$\sigma = r \cdot \min(\text{non-zero precipitation values}) \tag{14}$$

where $r = 0.1$ by default. This choice ensures that: (1) the perturbation magnitude is small relative to actual precipitation amounts, (2) the statistical properties of the original distribution are preserved, and (3) gradient optimization stability is enhanced without introducing significant bias.

For our South Australian precipitation data, this typically results in $\sigma \approx 0.01$ mm, which is negligible compared to meaningful precipitation events (typically > 1 mm) but sufficient to smooth the optimization landscape around zero values.

Solution 4: Weighted Tweedie Loss for Consistent Optimization

To maintain consistency between our attention mechanism and optimization objective, we design a weighted loss function that emphasizes non-zero events using the same weighting scheme as the attention computation. This unified weighting approach ensures that the model's learning objective aligns with its attention allocation strategy.

Theorem 5 (Convexity of the Weighted Tweedie Loss)**.** *The weighted Tweedie loss function*

$$L(\theta) = -\sum_{i=1}^{N} w_i \log f(y_i; \mu_i, \phi, \rho)$$

is convex with respect to the mean parameter μ under mild regularity conditions when $1 < \rho < 2$, which facilitates stable optimization using standard first-order methods.

Proof Sketch. For $1 < \rho < 2$, the Tweedie log-likelihood is concave with respect to the mean parameter μ. Specifically, its second derivative satisfies

$$\frac{\partial^2}{\partial \mu^2} \log f(y; \mu, \phi, \rho) = -\frac{(2-\rho)(1-\rho)}{\phi} \mu^{-\rho} < 0. \tag{15}$$

Therefore, the negative log-likelihood is convex in μ. Since the weights w_i are non-negative, the weighted sum preserves convexity. A detailed derivation is provided in Appendix A.5. □

The explicit form of our weighted Tweedie loss is:

$$L(\theta) = \sum_{i=1}^{N} \frac{w_i}{\phi} \left(\frac{y_i^{2-\rho}}{(1-\rho)(2-\rho)} - \frac{y_i \mu_i^{1-\rho}}{1-\rho} + \frac{\mu_i^{2-\rho}}{2-\rho} \right) \tag{16}$$

where the weights w_i follow the same fixed shared weight scheme used in attention:

$$w_i = \begin{cases} \alpha, & \text{if } y_i = 0 \\ \beta, & \text{if } y_i > 0 \end{cases} \tag{17}$$

This design choice ensures that: (1) the loss function emphasizes learning from non-zero precipitation events, (2) the attention mechanism and optimization objective are aligned,

and (3) the model maintains consistent focus throughout training.

The gradient of the weighted loss with respect to the mean parameter μ_i is:

$$\frac{\partial L}{\partial \mu_i} = \frac{w_i}{\phi}\left(\mu_i^{1-\rho} - y_i\mu_i^{-\rho}\right) \tag{18}$$

This gradient naturally balances the contribution of zero and non-zero samples according to their importance weights, leading to more stable and effective training on zero-inflated data.

Integration and Theoretical Guarantees

Our complete DET framework addresses all fundamental challenges identified in the theoretical analysis while providing rigorous guarantees for both optimization and generalization. The integration of weighted attention, Tweedie parameterization, Gaussian perturbation, and consistent loss weighting creates a unified architecture that maintains both statistical rigor and computational efficiency.

For generalization properties, our model satisfies the PAC-Bayes bound:

$$\mathbb{E}[L(\hat{\theta})] \leq \hat{L}(\hat{\theta}) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(\frac{1}{\delta})}{2n_{\text{eff}}}} \tag{19}$$

where the effective sample size is adjusted for zero-inflation:

$$n_{\text{eff}} = n \cdot \frac{2p_0p_1}{\alpha p_0 + \beta p_1} \tag{20}$$

This bound demonstrates that our weighting scheme improves generalization by balancing the effective contribution of zero and non-zero samples, preventing the model from overfitting to the abundant zero observations.

The method seamlessly integrates with nonstationary modeling capabilities by incorporating our weighted attention into the de-stationary attention mechanism of the Nonstationary Transformer:

$$\begin{aligned}\mathrm{Attn}(Q', K', V', \tau, \Delta, W) = \\ \mathrm{softmax}\left(\frac{\left(\tau Q' K'^T + 1\Delta^T\right) \odot W}{\sqrt{d_k}}\right) V'\end{aligned} \tag{21}$$

where τ and Δ are de-stationary factors derived from input statistics, and W contains our learned weights $w_{i,j}$. The element-wise multiplication $\odot$ allows both mechanisms to operate simultaneously, enabling the model to handle zero-inflation and nonstationarity concurrently.

The complete training procedure integrates all components: (1) apply Gaussian perturbation to zero-valued targets during preprocessing, (2) compute weighted attention scores using the fixed shared weight scheme, (3) predict Tweedie distribution parameters through separate MLPs, (4) calculate the weighted Tweedie loss emphasizing non-zero events, and (5) optimize using Adam with learning rate scheduling. The theoretical guarantees ensure stable convergence and optimal performance on zero-inflated precipitation forecasting tasks.

Experiments

Our study focuses on South Australia, a region known for its pronounced rainfall variability and recurrent droughts. The climate of South Australia is largely semi-arid to arid, characterized by significant fluctuations in rainfall driven by seasonal influences, including the southern jet stream and the influx of tropical moisture. Understanding these rainfall patterns is vital for managing agriculture, water resources, and climate adaptation strategies. This makes South Australia a compelling region for advancing and validating models designed to predict zero-inflated rainfall events.

Our dataset uses NCEP-NCAR Reanalysis 1[1], provided by the National Oceanic and Atmospheric Administration (NOAA). This data is available at a daily frequency, covering the period from 1948 to 2014, with a global resolution of 2.5 degrees in both the latitudinal and longitudinal directions, and has not been re-gridded. The nearest geographical point to South Australia was selected.

Data Preprocessing

We extracted and converted climate data from 1948 to 2014 for the South Australia region. The data, originally in NetCDF format, were transformed into CSV files for further processing. Our features include relevant meteorological variables, such as temperature, pressure, humidity, and wind speed, among others. Detailed data preprocessing procedures are provided in the Appendix.

Experimental Setting

We evaluate our method on NCEP-NCAR Reanalysis precipitation data from South Australia (1948-2014) with 80% zero values, using chronological train/validation/test splits of 70/15/15. Models use 96-step input sequences to predict 24, 48, 96, 192, 336, and 720 future steps. We compare against 9 baseline methods including Transformer variants, TSMixer, and statistical models (ZIP, Hurdle) using identical training configurations. More details can be found in the Appendix provided in the github repo https://github.com/Wentao-Gao/AAAI2026_Deep_Extreme_Transformer.

We carefully chose 7 well-acknowledged forecasting models and 2 zero-inflated models as our benchmark, including TSMixer (Chen et al. 2023), iTransformer (Liu et al. 2023), Transformer (Vaswani et al. 2017), PatchTST (Nie et al. 2023), TimeMixer (Wang et al. 2024), TimesNet (Wu et al. 2023), Nonstationary Transformer (Liu et al. 2022), Zero-Inflated Poisson (ZIP), and Hurdle Model to ensure a comprehensive evaluation of our approach.

Main Results

As shown in Table 1, our experimental results demonstrate that DET significantly outperforms all baseline methods across forecasting horizons. For 24-step forecasting, DET achieves the best performance with MSE of 0.512 and MAE of 0.212. Compared to the strongest deep learning baselines,

[1]https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html

Output Length		DeepExtreme Transformer (Ours)	TSMixer (2023)	iTransformer (2024)	Transformer (2017)	PatchTST (2023)	TimeMixer (2024)	TimesNet (2023)	Nonstationary Transformer (2022)	ZIP (1992)	Hurdle (1986)
24	**MSE**	**0.5120**	0.9458	0.9018	0.8723	0.9065	0.9089	0.6520	0.7200	1.5620	1.8725
	MAE	**0.2120**	0.4010	0.3256	0.3157	0.3251	0.3140	0.4291	0.2975	0.5612	0.6250
48	**MSE**	**0.5180**	0.9491	0.9082	0.8641	0.9118	0.9127	0.6620	0.7350	1.6031	1.8952
	MAE	**0.2140**	0.4084	0.3171	0.3312	0.3243	0.3211	0.4337	0.3117	0.5763	0.6425
96	**MSE**	**0.5320**	0.9546	0.9220	0.8716	0.9185	0.9205	0.6776	0.7600	1.7025	1.9331
	MAE	**0.2210**	0.4098	0.3299	0.2973	0.3067	0.3156	0.4455	0.3078	0.5889	0.6520
192	**MSE**	**0.5450**	0.9671	0.9292	0.8764	0.9200	0.9252	0.7450	0.7800	1.7501	1.9803
	MAE	**0.2280**	0.4053	0.3339	0.3090	0.3200	0.3220	0.4904	0.2984	0.6025	0.6624
336	**MSE**	**0.5580**	0.9734	0.9215	0.8777	0.9250	0.9267	0.7500	0.8000	1.8457	2.0310
	MAE	**0.2350**	0.4253	0.3217	0.3048	0.3150	0.3235	0.4950	0.3026	0.6205	0.6742
720	**MSE**	**0.5800**	0.9733	0.9179	0.8655	0.9300	0.9490	0.7600	0.8200	1.9123	2.1057
	MAE	**0.2450**	0.4282	0.3150	0.3050	0.3200	1.8307	0.5000	0.3080	0.6358	0.6820

Table 1: Long-term Forecast Results (Including ZIP and Hurdle Models)

Design		48		96		192		336	
		MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Full Model	Weighted Attention + Tweedie Loss	**0.5180**	**0.2140**	**0.5320**	**0.2210**	**0.5450**	**0.2280**	**0.5580**	**0.2350**
Replace	Tweedie Loss → Standard Loss	0.620	0.265	0.650	0.275	0.680	0.285	0.700	0.295
	Weighted Attention → Standard Attention	0.610	0.260	0.640	0.270	0.670	0.280	0.690	0.290
	Tweedie Mapping → Standard Regression	0.605	0.258	0.635	0.268	0.665	0.278	0.685	0.288
w/o	w/o Weighted Attention	0.630	0.270	0.660	0.280	0.690	0.290	0.720	0.300
	w/o Gaussian Perturbation	0.635	0.275	0.665	0.285	0.695	0.295	0.725	0.305

Table 2: Ablation Study Results on Zero-Inflated Precipitation Data

DET shows consistent improvements: 21.5% MSE reduction over TimesNet (0.512 vs 0.652) and 28.7% MAE reduction over Nonstationary Transformer (0.212 vs 0.298).

The performance advantage remains consistent across longer horizons, with MSE improvements of 21.8% (48-step), 21.5% (96-step), and 26.8% (192-step) compared to TimesNet. For MAE, improvements range from 23.6% to 31.3% compared to Nonstationary Transformer across different horizons. Notably, statistical models designed specifically for zero-inflated data (ZIP and Hurdle) perform significantly worse, with MSE values 2-3 times higher than DET, highlighting the critical importance of combining temporal modeling with zero-inflation handling.

Previous studies indicate that extending the lookback window does not always improve forecasting accuracy and may even degrade performance due to attention distraction (Nie et al. 2023; Zeng et al. 2022). Hence, we fix the lookback length at 96 timesteps to balance accuracy and efficiency. Unlike other Transformer variants affected by attention dilution in zero-inflated data, DET captures key patterns through its extreme attention mechanism, achieving state-of-the-art performance in zero-inflated time series forecasting.

Ablation Study

To validate the contributions of each component in our Deep Extreme Transformer, we conducted systematic ablation studies with both replacement and removal experiments as shown in Table 2. The results demonstrate that each proposed component is essential for optimal performance. For 96-step forecasting, replacing the Tweedie loss with standard MSE loss degrades MSE from 0.532 to 0.650 (22.2% performance loss), confirming that the Tweedie distribution is critical for modeling zero-inflated continuous data. Substituting weighted attention with standard attention increases MSE to 0.640 (20.3% worse), demonstrating that our weighting mechanism successfully addresses attention dilution in zero-inflated sequences. The removal experiments reveal even larger performance degradations, with removing Gaussian perturbation causing the most severe impact (MSE increases to 0.665, 25.0% worse performance) and removing weighted attention leading to substantial degradation (MSE: 0.660, 24.1% performance loss). Consistently across all prediction horizons, component removal causes larger performance drops than component replacement, and all components contribute meaningfully to handling zero-inflated precipitation forecasting challenges.

Conclusion

We introduced the DET, a unified architecture for zero-inflated rainfall prediction that integrates Tweedie distribution modeling, weighted attention, and Gaussian perturbation. DET consistently surpasses state-of-the-art methods across horizons up to 720 steps, achieving an MSE of 0.512 and MAE of 0.212 at 24-step forecasts. In future work, we aim to incorporate causal representations (Xu et al. 2024; Cheng et al. 2023, 2024) to disentangle latent drivers of extreme events and improve interpretability, extending DET toward causality-aware zero-inflated forecasting. This direction will enable deeper insights into climate dynamics and broader applications across complex, data-sparse temporal domains.

References

Beveridge, M.; Goldstein, Z.; and Chung, H. C. 2024. A Comparison of Zero-Inflated Models for Modern Biomedical Data. *arXiv preprint arXiv:2411.12086*.

Billingsley, P. 2013. *Convergence of probability measures*. John Wiley & Sons.

Brockwell, P. J.; and Davis, R. A. 2002. *Introduction to time series and forecasting*. Springer.

Chen, S.-A.; Li, C.-L.; Yoder, N.; Arik, S. O.; and Pfister, T. 2023. TSMixer: An All-MLP Architecture for Time Series Forecasting. arXiv:2303.06053.

Cheng, D.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; Gao, W.; and Le, T. D. 2024. Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11480–11488.

Cheng, D.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; and Le, T. D. 2023. Causal inference with conditional instruments using deep generative models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 7122–7130.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Coe, R.; and Stern, R. 1982. Fitting models to daily rainfall data. *Journal of Applied Meteorology (1962-1982)*, 1024–1031.

Cox, D. R. 1955. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2): 129–157.

Gao, W.; Li, J.; Cheng, D.; Liu, L.; Liu, J.; Le, T. D.; Du, X.; Chen, X.; Zhao, Y.; and Chen, Y. 2024a. A deconfounding approach to climate model bias correction. *arXiv preprint arXiv:2408.12063*.

Gao, W.; Li, J.; Liu, L.; Le, T. D.; Chen, X.; Du, X.; Liu, J.; Zhao, Y.; and Chen, Y. 2025. From Noise to Precision: A Diffusion-Driven Approach to Zero-Inflated Precipitation Prediction. *arXiv preprint arXiv:2509.10501*.

Gao, W.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; Le, T. D.; Cheng, D.; Zhao, Y.; and Chen, Y. 2024b. TSI: A Multi-view Representation Learning Approach for Time Series Forecasting. In *Australasian Joint Conference on Artificial Intelligence*, 291–302. Springer.

Han, Z.; Zhao, J.; Leung, H.; Ma, K. F.; and Wang, W. 2019. A review of deep learning models for time series prediction. *IEEE Sensors Journal*, 21(6): 7833–7848.

Kitaev, N.; Kaiser, Ł.; and Levskaya, A. 2020. Reformer: The Efficient Transformer. arXiv:2001.04451.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1): 1–14.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35: 9881–9893.

Maiti, R.; Biswas, A.; Guha, A.; and Ong, S. H. 2014. Modelling and coherent forecasting of zero-inflated count time series. *Statistical Modelling*, 14(5): 375–398.

Milly, P. C.; Betancourt, J.; Falkenmark, M.; Hirsch, R. M.; Kundzewicz, Z. W.; Lettenmaier, D. P.; and Stouffer, R. J. 2008. Stationarity is dead: Whither water management? *Science*, 319(5863): 573–574.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3): 341–365.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730.

Norel, M.; Kałczyński, M.; Pińskwar, I.; Krawiec, K.; and Kundzewicz, Z. W. 2021. Climate variability indices—a guided tour. *Geosciences*, 11(3): 128.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; and Prabhat, F. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8: 842–866.

Schepen, A.; Wang, Q.; and Robertson, D. E. 2012. Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *Journal of Geophysical Research: Atmospheres*, 117(D20).

Shi, Y.; Dai, W.; and Long, W. 2021. A new deep learning-based zero-inflated duration model for financial data irregularly spaced in time. *Frontiers in Physics*, 9: 651528.

Vandal, T.; Kodra, E.; Ganguly, S.; Michaelis, A.; Nemani, R.; and Ganguly, A. R. 2017. DeepSD: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1663–1672.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. arXiv:2405.14616.

Wilby, R. L.; Wigley, T.; Conway, D.; Jones, P.; Hewitson, B.; Main, J.; and Wilks, D. 1998. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, 34(11): 2995–3008.

Wilks, D. S. 2011. *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

Wood, A. W.; Maurer, E. P.; Kumar, A.; and Lettenmaier, D. P. 2002. Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, 107(D20): ACL–6.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. arXiv:2210.02186.

Xu, Z.; Cheng, D.; Li, J.; Liu, J.; Liu, L.; and Yu, K. 2024. Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder. In *The Twelfth International Conference on Learning Representations*.

Xu, Z.; Kandanaarachchi, S.; Ong, C. S.; and Ntoutsi, E. 2025. Fairness evaluation with item response theory. In *Proceedings of the ACM on Web Conference 2025*, 2276–2288.

Yu, W.; Li, J.; Gao, W.; Zhuang, N.; Li, W.; and Du, S. 2025. G-GLformer: Transformer with GRU Embedding and Global-Local Attention for Multivariate Time Series Forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 41–56. Springer.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2022. Are Transformers Effective for Time Series Forecasting? arXiv:2205.13504.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zucchini, W.; and MacDonald, I. L. 2009. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.