

Probabilistic Flood Catastrophe Modeling for Urban Wastewater Infrastructure Using Generative Tail Risk and Systemic Climate Vulnerability

Sandeep Sahu

Climate Risk & Geospatial AI

15 January 2026

Abstract

Urban wastewater infrastructure is increasingly exposed to extreme flood risk driven by intensifying precipitation, urban expansion, and non-stationary climate dynamics. Deterministic flood assessments fail to represent tail risk, event clustering, and infrastructure-scale financial and social losses critical for resilience planning, insurance, and policy decisions.

This study presents a fully probabilistic, facility-level flood catastrophe risk modeling framework applied to the Musi River sub-basin, India. The framework integrates observed rainfall, reanalysis data, Extreme Value Theory (EVT), Monte Carlo loss simulation, deep generative modeling using Variational Autoencoders (VAE), and spatially explicit hazard–exposure–adaptive capacity indices.

The resulting system quantifies extreme rainfall hazard, infrastructure damage, financial tail losses, social impacts, climate stress amplification, and composite climate vulnerability in a manner aligned with contemporary catastrophe modeling and reinsurance analytics practice.

1 Problem Context and Motivation

Urban wastewater systems are critical lifeline infrastructure whose failure during floods can trigger cascading public health, environmental, and economic impacts. In rapidly urbanizing river basins such as the Musi, flood risk is governed not only by extreme rainfall but also by terrain configuration, drainage connectivity, population exposure, and institutional capacity.

Conventional flood risk assessments typically rely on deterministic inundation maps or single return-period design events. Such approaches fail to:

- Quantify probabilistic tail risk and loss exceedance behavior,
- Capture multi-event years and loss aggregation,
- Represent non-parametric behavior of extreme rainfall tails,
- Integrate socio-economic exposure and adaptive capacity.

This work addresses these limitations through a catastrophe modeling paradigm that combines statistical extreme value theory, stochastic simulation, deep generative modeling, and spatial risk indexing.

2 Study Area and Asset Inventory

The Musi River sub-basin, located in Telangana, India, is selected as the study domain. The analysis focuses on sixteen wastewater-related facilities, including wastewater treatment plants, landfills, transfer stations, and recycling facilities.

Facilities are represented as point assets. Terrain attributes (elevation, slope, distance to drainage) and population exposure metrics are derived through geospatial preprocessing pipelines and associated with each facility.

3 Overall Modeling Framework

Figure 1 summarizes the complete modeling pipeline implemented in this study.

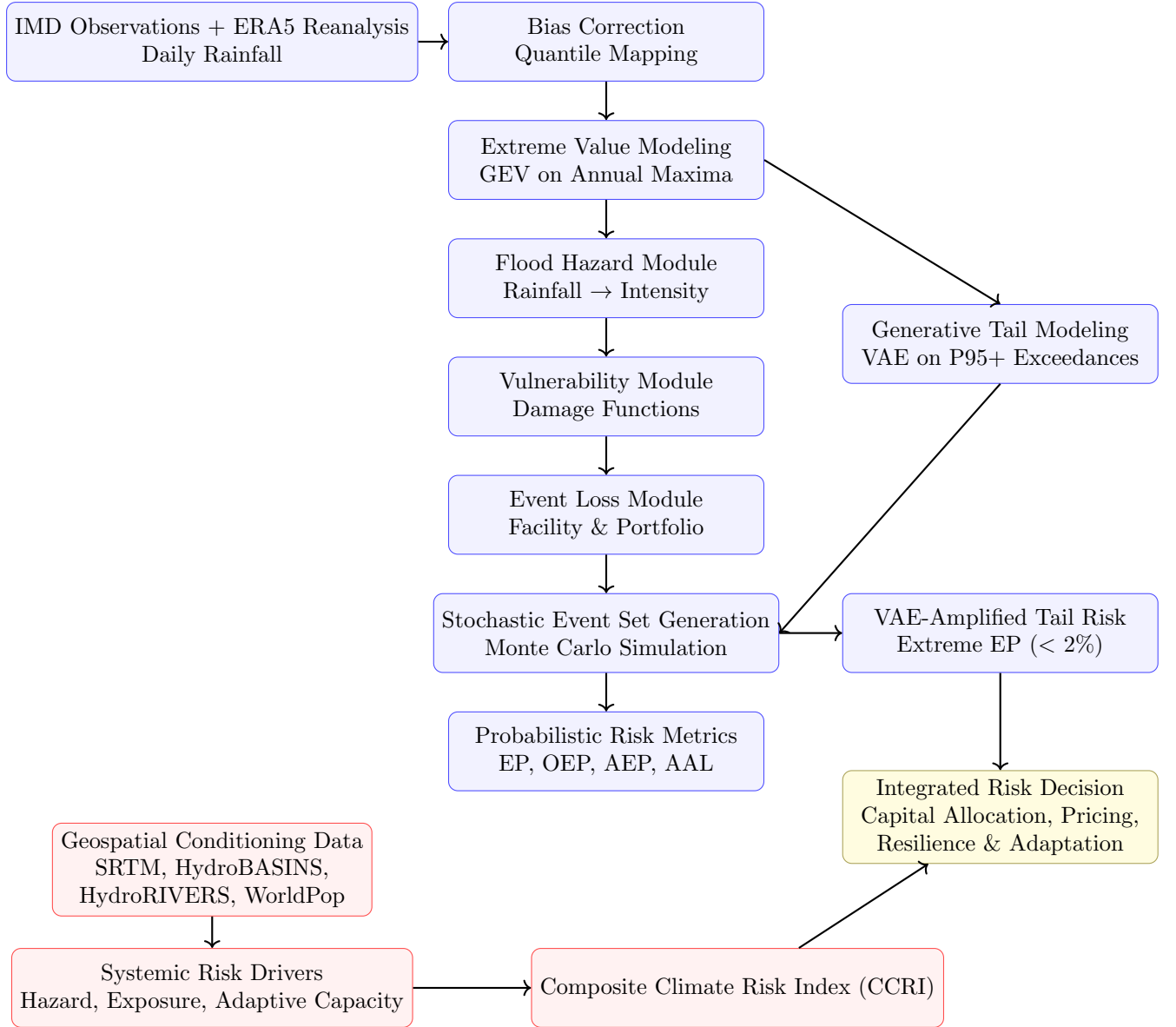


Figure 1: Probabilistic flood catastrophe modeling framework combining EVT-conditioned stochastic loss modeling and generative tail amplification for rare events ($EP < 2\%$) (blue) with a parallel systemic climate risk pipeline (red) producing a Composite Climate Risk Index (CCRI). Both probabilistic tail risk and structural vulnerability inform an integrated downstream risk decision layer.

4 Data Sources and Preprocessing

4.1 Observed Rainfall (IMD)

Daily gridded rainfall from the India Meteorological Department (IMD) is used as the observational reference dataset. IMD rainfall is spatially clipped to the Musi sub-basin and aggregated to basin-mean daily rainfall for the period 2010–2023.

4.2 Reanalysis Rainfall (ERA5)

ERA5 daily total precipitation provides a spatially and temporally continuous rainfall record. ERA5 rainfall is clipped to the Musi sub-basin and aggregated to basin-mean daily rainfall for the same period.

4.3 Bias Correction

ERA5 rainfall is bias-corrected against IMD observations using non-parametric quantile mapping:

$$R^{BC} = F_{IMD}^{-1}(F_{ERA5}(R))$$

Negative rainfall values are clipped to zero.

5 Extreme Rainfall Modeling using Extreme Value Theory

Extreme rainfall governs flood generation and infrastructure stress in the Musi basin. To statistically characterize rare but high-impact rainfall events, Extreme Value Theory (EVT) is employed using the Annual Maximum Series (AMS) approach.

Bias-corrected basin-mean daily rainfall is first aggregated by calendar year, and the maximum daily rainfall from each year is retained. Let $\{x_1, x_2, \dots, x_n\}$ denote the annual maxima over n years.

The AMS is modeled using the Generalized Extreme Value (GEV) distribution, whose cumulative distribution function is given by:

$$G(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \quad 1 + \xi \left(\frac{x - \mu}{\sigma} \right) > 0$$

where:

- μ is the location parameter (controls central tendency),
- $\sigma > 0$ is the scale parameter (controls dispersion),
- ξ is the shape parameter (controls tail heaviness).

The parameters are estimated via maximum likelihood estimation applied to the AMS, yielding:

$$\mu = 48.47 \text{ mm}, \quad \sigma = 15.35 \text{ mm}, \quad \xi = 0.094$$

Return levels corresponding to a return period T (years) are computed as:

$$R_T = G^{-1}(1 - 1/T)$$

The 100-year return level,

$$R_{100} = 136.87 \text{ mm},$$

is used throughout the framework as a normalization anchor for event-level flood hazard scaling.

6 Flood Hazard Modeling

Flood hazard is modeled in two stages: (i) event-level hazard driven by rainfall severity, and (ii) facility-level modulation driven by local bio-physical susceptibility.

6.1 Event-Level Hazard

For a simulated flood event e with rainfall intensity R_e , a normalized hazard index is defined as:

$$H_e = \min\left(\frac{R_e}{R_{100}}, 1\right)$$

This formulation expresses event severity relative to the 100-year rainfall benchmark while bounding hazard within the unit interval $[0, 1]$.

6.2 Facility Flood Susceptibility

Each facility i is assigned a static flood susceptibility index H_i derived from terrain and drainage characteristics. Three normalized bio-physical components are used:

- Normalized distance to drainage (d_i^*),
- Elevation percentile ($p_{elev,i}$),
- Slope percentile ($p_{slope,i}$).

Lower elevation, flatter slopes, and proximity to drainage increase flood susceptibility. The composite susceptibility index is defined as:

$$H_i = 0.40(1 - d_i^*) + 0.35(1 - p_{elev,i}) + 0.25(1 - p_{slope,i})$$

The weights reflect the dominant influence of drainage proximity, followed by elevation and slope.

6.3 Facility Hazard Intensity

Facility-level hazard for event e is obtained by multiplicative scaling:

$$H_{i,e} = H_e \cdot H_i$$

This formulation ensures that extreme rainfall only translates into high facility hazard where local conditions permit flood accumulation.

7 Damage and Asset Modeling

Infrastructure damage is modeled as a nonlinear function of hazard intensity, reflecting threshold effects and rapid damage escalation under severe flooding.

7.1 Damage Functions

For facility i during event e , the damage ratio is defined as:

$$D_{i,e} = \min(H_{i,e}^{\alpha_i}, 1)$$

where $\alpha_i > 1$ is a facility-type-specific vulnerability exponent controlling damage convexity.

Facility Type	Vulnerability Exponent α_i
Wastewater Plant	1.3
Transfer Station	1.6
Recycling Facility	1.8
Landfill	2.0
Other	2.2

Higher α_i values represent more fragile or exposure-sensitive facilities.

7.2 Asset Values

Facility replacement values V_i are assigned based on infrastructure criticality and relative capital intensity:

Facility Type	Asset Value
Wastewater Plant	5×10^7
Landfill	2×10^7
Transfer Station	1.5×10^7
Recycling Facility	1×10^7
Other	1×10^7

The absolute currency unit is immaterial; relative scaling governs loss behavior.

7.3 Event-Level Loss

Financial loss for facility i during event e is computed as:

$$L_{i,e} = V_i \cdot D_{i,e}$$

8 Monte Carlo Loss Simulation

To represent interannual variability, event clustering, and tail risk, losses are simulated using Monte Carlo techniques.

8.1 Event Frequency

The number of flood events in year y is modeled as a Poisson process:

$$N_y \sim \text{Poisson}(\lambda), \quad \lambda = 0.4$$

This reflects intermittent but recurring flood-generating rainfall events.

8.2 Annual Loss Aggregation

Annual portfolio loss is computed as:

$$L_y = \sum_{e=1}^{N_y} \sum_i L_{i,e}$$

A total of

$$N = 10,000 \text{ simulation years}$$

are generated to construct empirical loss distributions.

9 Risk Metrics

The simulated annual loss distribution is summarized using standard catastrophe risk metrics.

Occurrence Exceedance Probability (OEP):

$$\text{OEP}(x) = P\left(\max_e L_e > x\right)$$

which captures single-event dominance.

Aggregate Exceedance Probability (AEP):

$$\text{AEP}(x) = P(L_y > x)$$

which captures multi-event accumulation within a year.

Average Annual Loss (AAL):

$$\text{AAL} = \int_0^\infty P(L > x) dx$$

representing the expected long-term annual loss.

10 Casualty Modeling

Social impacts are modeled by estimating expected casualties as a function of hazard intensity and population exposure.

10.1 Casualty Fraction

The fraction of exposed population affected is modeled as:

$$f(h) = \min(ch^\gamma, 0.2)$$

where c and γ are facility-type-specific parameters controlling severity and nonlinearity.

10.2 Expected Casualties

Expected casualties for facility i during event e are:

$$C_{i,e} = f(H_{i,e})P_i$$

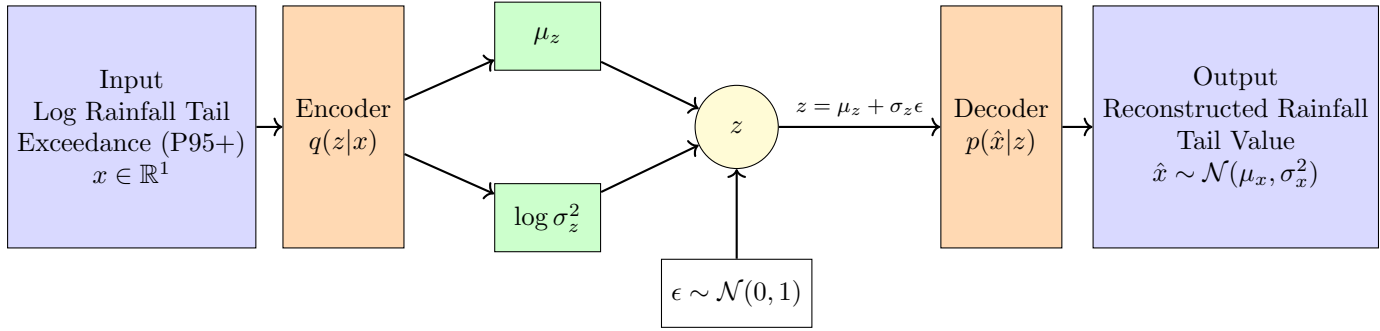
where P_i denotes population exposure near the facility.

10.3 Expected Annual Casualties

Social risk is summarized via:

$$EAC = \int_0^\infty P(C > x) dx$$

11 VAE-Based Tail Modeling



$$\begin{aligned} & \text{Maximize (ELBO)} \\ \mathcal{L}_{\text{ELBO}} = & \underbrace{\frac{1}{2} \left[\log \sigma_x^2 + \frac{(x - \hat{x})^2}{\sigma_x^2} \right]}_{\text{Gaussian Reconstruction NLL}} + \underbrace{\frac{1}{2} (\mu_z^2 + \sigma_z^2 - \log \sigma_z^2 - 1)}_{\text{KL}(q(z|x) \| p(z))} \end{aligned}$$

Figure 2: Variational autoencoder (VAE) used for rainfall tail enrichment. The input x represents log rainfall tail exceedances above the 95th percentile (P95+). The encoder maps x to a latent Gaussian representation, from which samples are drawn using the reparameterization trick. The decoder reconstructs the rainfall tail value \hat{x} , and training maximizes the evidence lower bound (ELBO), consistent with the implemented architecture and loss formulation.

While EVT captures parametric tail behavior, additional flexibility is introduced using deep generative modeling.

11.1 Tail Extraction

Let R denote daily rainfall. Extreme rainfall values are isolated by extracting exceedances above the empirical 95th percentile threshold q_{95} :

$$X = R - q_{95}, \quad R > q_{95}$$

To stabilize variance and improve numerical conditioning in the tail regime, a logarithmic

transformation is applied:

$$x = \log(X)$$

Here, $x \in \mathbb{R}^1$ represents the *log rainfall tail exceedance (P95+)*, which forms the input to the generative tail model.

11.2 Variational Autoencoder for Tail Modeling

A Variational Autoencoder (VAE) is trained on the log-transformed tail exceedances x to learn a smooth latent representation of extreme rainfall behavior. The encoder defines an approximate posterior:

$$q(z | x) = \mathcal{N}(\mu_z(x), \sigma_z^2(x)),$$

where $z \in \mathbb{R}^1$ is a low-dimensional latent variable. Sampling is performed using the reparameterization trick:

$$z = \mu_z + \sigma_z \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1).$$

The decoder parameterizes the conditional distribution of the reconstructed rainfall tail value \hat{x} :

$$p(\hat{x} | z) = \mathcal{N}(\mu_x(z), \sigma_x^2(z)).$$

Model training maximizes the Evidence Lower Bound (ELBO), given by:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\frac{1}{2} \left[\log \sigma_x^2 + \frac{(x - \hat{x})^2}{\sigma_x^2} \right]}_{\text{Gaussian Reconstruction NLL}} + \underbrace{\frac{1}{2} (\mu_z^2 + \sigma_z^2 - \log \sigma_z^2 - 1)}_{\text{KL}(q(z|x) \| p(z))}$$

which corresponds to a Gaussian negative log-likelihood reconstruction term and a Kullback–Leibler divergence regularizer, consistent with the implemented loss function.

The VAE is applied exclusively within the extreme-value regime and does not replace classical EVT modeling, but instead enriches the representation of rainfall tail variability.

11.3 Tail Injection in Monte Carlo Simulation

During stochastic event generation, rainfall intensity samples are drawn from a mixture of classical EVT-based tail estimates and the VAE-generated tail distribution. Specifically, with probability

$$P_{\text{tail}} = 0.05,$$

samples are drawn from the VAE decoder distribution $p(\hat{x} | z)$, while with probability $1 - P_{\text{tail}}$, samples are drawn from the EVT-based tail model.

This controlled tail injection mechanism enriches rare-event severity while preserving the statistical structure of the primary EVT model, thereby improving representation of extreme loss behavior without distorting bulk rainfall characteristics.

12 Composite Climate Risk Index

Beyond probabilistic financial loss, systemic climate vulnerability of wastewater infrastructure is quantified using a Composite Climate Risk Index (CCRI). The CCRI is designed to capture long-term, non-financial climate stress arising from the interaction of physical susceptibility, socio-economic exposure, and adaptive capacity. All components are constructed as normalized, relative indices to enable comparison across facilities within the Musi sub-basin.

12.1 Hazard Component

The hazard term H_i represents facility-level flood susceptibility driven by static hydro-terrain controls, independent of event magnitude. It is derived from a weighted combination of normalized biophysical indicators:

$$H_i = w_d D_i + w_e ELEV_i + w_s SLP_i$$

where:

- D_i is normalized proximity to drainage (inverse distance to nearest river or channel),
- $ELEV_i$ is the percentile-normalized elevation (lower elevations imply higher hazard),
- SLP_i is the percentile-normalized slope (lower slopes imply higher flood accumulation potential),

with weights $w_d + w_e + w_s = 1$. These variables collectively represent terrain-controlled flood accumulation and persistence rather than event-driven severity.

12.2 Socio-Economic Exposure

Exposure captures the potential societal and service disruption associated with facility failure. It is defined as:

$$E_i = \text{norm}(P_i \cdot S_i)$$

where:

- P_i is population exposure derived from spatial population density in the facility service area,
- S_i is service criticality, reflecting the functional importance of the facility type (e.g., wastewater treatment plants receiving higher weights than transfer stations).

The exposure term is normalized across all facilities to ensure comparability and to prevent domination by absolute population size.

12.3 Adaptive Capacity

Adaptive capacity represents the relative ability of a facility to cope with, respond to, and recover from flood impacts. It is constructed as a weighted composite of proxy indicators:

$$C_i = 0.35I_i + 0.30R_i + 0.20A_i + 0.15(1 - P_i)$$

where:

- I_i denotes institutional and governance capacity, approximated using facility type and administrative context,
- R_i represents physical robustness and redundancy, reflecting infrastructure scale and engineering resilience,
- A_i captures accessibility and response reach, proxied by connectivity to transport and emergency access,
- $(1 - P_i)$ represents population buffering, accounting for reduced recovery capacity under high population pressure.

All components are normalized to $[0, 1]$, with higher values indicating greater adaptive capacity. The weights reflect the relative importance of institutional and physical factors observed in urban infrastructure resilience literature.

12.4 Composite Climate Risk Index

The final Composite Climate Risk Index is defined multiplicatively as:

$$\text{CCRI}_i = H_i \cdot E_i \cdot (1 - C_i)$$

This formulation ensures that high systemic climate risk emerges only where physical flood susceptibility and socio-economic exposure coincide with limited adaptive capacity. The use of $(1 - C_i)$ explicitly models adaptive capacity as a mitigating factor rather than a risk amplifier.

Higher CCRI values therefore identify facilities that are simultaneously flood-prone, socially consequential, and institutionally constrained, representing priority assets for climate adaptation and resilience planning.

13 Results

- Flood risk and climate vulnerability are spatially concentrated along the Musi River corridor, where low-relief terrain, drainage convergence, and dense population exposure jointly elevate systemic risk to wastewater infrastructure.
- Composite Climate Risk is not driven by flood hazard alone: facilities with moderate biophysical susceptibility emerge as high-risk assets when socio-economic exposure and low adaptive capacity are considered, confirming the necessity of integrated hazard–exposure–capacity modeling.
- EVT-only flood modeling underestimates extreme loss uncertainty, as return-level-based loss curves are limited to sparse extremes and do not represent annual loss aggregation or compound event behavior.

- Monte Carlo simulation substantially improves loss estimation by aggregating stochastic event frequency over 10,000 synthetic years, producing stable EP curves and revealing materially higher losses at low exceedance probabilities relative to EVT-only estimates.
- Deep generative tail amplification using a VAE selectively thickens the extreme loss tail without altering mid-range or average risk, with divergence from Monte Carlo estimates occurring only at very low exceedance probabilities ($EP < 10^{-3}$).
- Climate stress manifests primarily through non-linear amplification of extreme tail losses, rather than shifts in typical losses or event frequency, with a small subset of structurally tail-sensitive facilities dominating portfolio losses under severe stress scenarios.

14 Interpretation of Risk Perspectives

The proposed framework explicitly distinguishes between two complementary but conceptually distinct risk perspectives:

- **Financial Tail Risk**, and
- **Composite Climate Risk**.

Financial Tail Risk represents the probability and magnitude of low-frequency, high-impact loss events that dominate capital requirements. This perspective is quantified through stochastic loss simulation, exceedance probability (EP) curves, Average Annual Loss (AAL), and Probable Maximum Loss (PML). It captures the behavior of extreme rainfall-driven flood events, event clustering, and nonlinear damage amplification, and is directly aligned with insurance, reinsurance, and financial stress-testing applications.

In contrast, **Composite Climate Risk** represents systemic vulnerability at the facility level, integrating bio-physical flood susceptibility, socio-economic exposure, and adaptive capacity. This index-based perspective does not aim to estimate monetary loss directly; rather, it identifies assets that are structurally, socially, and institutionally fragile under climate stress. Composite Climate Risk therefore reflects chronic vulnerability and resilience gaps that may not be immediately apparent from financial loss metrics alone.

The joint interpretation of these two perspectives is essential. Facilities exhibiting moderate financial tail losses may still represent critical climate risks due to high population exposure or low adaptive capacity, while assets with large modeled losses may be well-managed and resilient from an operational standpoint. Together, these perspectives enable decision-makers to balance capital risk, public safety, and long-term resilience.

15 Applications and Limitations

The modeling framework is designed to support a broad range of applied use cases, including:

- Climate risk and resilience consulting for urban infrastructure systems,
- Insurance and reinsurance portfolio risk assessment,

- Infrastructure investment prioritization and capital planning,
- Climate stress testing and scenario analysis for regulatory reporting,
- Policy-driven risk classification and adaptation planning.

Despite its comprehensive scope, the framework involves several simplifying assumptions. Vulnerability and casualty parameters are stylized and may not fully capture facility-specific engineering characteristics. Population exposure is treated as static and does not account for temporal mobility or emergency response dynamics. Hydrodynamic flood routing, infrastructure interdependencies, and compound hazards such as flooding combined with power outages are not explicitly modeled. Additionally, the generative rainfall model is unconditional and does not incorporate explicit climate covariates.

These limitations reflect deliberate trade-offs between model interpretability, computational tractability, and data availability, and they provide clear directions for future refinement rather than undermining the validity of the present analysis.

16 Discussion

This study demonstrates that flood-related climate risk to urban wastewater infrastructure arises from the interaction of two fundamentally distinct dimensions: long-term systemic vulnerability shaped by terrain, exposure, and adaptive capacity, and short-term probabilistic financial risk governed by rare, high-consequence flood extremes.

The systemic climate vulnerability assessment highlights that infrastructure risk cannot be inferred from flood hazard alone. Facilities situated in similar hydro-terrain conditions exhibit markedly different risk profiles depending on population dependency and coping capacity. Adaptive capacity emerges as a critical moderating factor, preventing some high-hazard locations from becoming systemically critical while exposing others as priority assets despite moderate physical susceptibility.

From a catastrophe modeling perspective, EVT provides a necessary physical foundation for representing extreme rainfall but is inherently limited by data scarcity in the tail and its focus on individual event severity. Monte Carlo simulation resolves these limitations by embedding EVT within a stochastic loss aggregation framework, enabling estimation of annual exceedance probabilities and portfolio-scale loss behavior.

However, Monte Carlo simulation remains bounded by the parametric tail assumptions inherited from EVT. The introduction of VAE-based tail amplification addresses this limitation by learning a latent representation of rainfall excess behavior, enabling controlled exploration of rare and compound tail realizations beyond the support of observed extremes. Importantly, the VAE does not inflate typical or moderate losses; its impact is confined to the extreme tail, where uncertainty dominates risk.

Climate stress testing results indicate that climate change effects are better conceptualized as non-stationary tail thickening rather than uniform risk escalation. Moderate losses remain stable across stress scenarios, while extreme losses increase disproportionately under stronger

tail amplification. At the facility level, this amplification is highly uneven, concentrating risk in a small number of structurally sensitive assets that dominate portfolio tail losses.

The joint interpretation of VAE-amplified extreme tail losses and composite climate vulnerability reveals a systematic divergence between financially catastrophic assets and systemically vulnerable infrastructure. This divergence underscores the inadequacy of single-metric risk rankings and motivates differentiated intervention strategies aligned with the dominant risk mechanism.

17 Conclusion

This study presents an integrated climate–catastrophe risk modeling framework that unifies systemic climate vulnerability assessment with probabilistic flood loss modeling enhanced by deep generative tail amplification. Applied to urban wastewater infrastructure in the Musi River sub-basin, the framework demonstrates that climate risk is neither purely hydrological nor purely financial, but emerges from the interaction of extreme hazard tails, socio-economic exposure, and adaptive capacity.

By combining Extreme Value Theory, Monte Carlo catastrophe simulation, and Variational Autoencoder-based tail enrichment, the framework advances beyond conventional flood risk assessments and provides a credible, tail-sensitive basis for climate stress testing. Results show that average losses and standard return periods are insufficient to reveal climate-driven vulnerability; instead, rare, amplified tail events govern capital stress, reinsurance relevance, and resilience prioritization.

The integrated analysis enables asset-specific risk classification, distinguishing between financially critical, systemically critical, dual-risk, and low-priority infrastructure. This separation supports targeted intervention strategies ranging from capital protection and risk transfer to governance reform and adaptive capacity enhancement.

Overall, the study demonstrates how modern Climate AI methods can be rigorously integrated with catastrophe risk modeling to deliver interpretable, decision-relevant insights for infrastructure planning, financial risk management, and climate resilience under intensifying hydroclimatic extremes.