School of Engineering and Applied Science (SEAS), Ahmedabad University

CSE 400: Fundamentals of Probability in Computing

Milestone-2 Scribe

S2_G11_CLI

Sr.No	Enrollment No	Name	Email Id
1.	AU2440247	Jenil Modi	jenil.m1@ahduni.edu.in
2.	AU2440214	Nikkesh Parekh	nikkesh.p@ahduni.edu.in
3.	AU2420114	Heer Patel	heer.p1@ahduni.edu.in
4.	AU2440097	Chaitanya Jammula	jammula.m@ahduni.edu.in
5.	AU2440081	Akshita Muchhal	akshita.m2@ahduni.edu.in

Problem Statement :
Deterministic extrapolation of atmospheric conditions fails rapidly due to the chaotic nature of fluid dynamics. Specifically, quantitative precipitation nowcasting (QPN) suffers from unbounded error growth when modeled deterministically. The issue is to construct a rigorous **Stochastic State-Space Model** that captures the spatiotemporal evolution of rainfall. By isolating the deterministic advection component from the stochastic growth/decay component, we aim to generate continuous probability density functions (PDFs) of future weather states, allowing for robust uncertainty quantification and risk analysis for extreme weather events.

Question-1: Project System and Objective

Solution:

1.1 The Probabilistic Problem Formulation

We examine the temporal evolution of precipitation fields as a stochastic process on a defined probability space $(\Omega_{prob}, \mathcal{F},)$. Let $\Omega \subset^2$ denote the 2D spatial spatial domain (a geographical grid) and let $\mathcal{T} = [0, T]$ represent the continuous time horizon.

The deterministic approach to fluid dynamics attempts to solve the Navier-Stokes equations directly. However, due to sub-grid scale unobservability and the chaotic divergence inherent to atmospheric flows, a precise point-mass prediction $\hat{y}$ of the future state becomes mathematically ill-posed for lead times $\tau > 0$. Therefore, the problem is fundamentally probabilistic: we must map current, noisy observations to a predictive posterior probability measure over the future state space.

1.2 System Objective in the State-Space Framework

In Milestone 2, our objective is to operationalize a **Stochastic State-Space formulation** constructed strictly around three interacting random fields. Let $Z_{1:t}$ represent the sequence of discrete radar observations up to time t. We seek to formulate the state transition dynamics such that we can continuously sample from the posterior predictive distribution:

$$f_{R_{t+\tau}|Z_{1:t}}(r \mid z) \quad \text{for any lead time } \tau > 0$$

By generating an ensemble of M independent and identically distributed (i.i.d.) realizations from this distribution, we can evaluate the expected intensity and bound the variance:

$$\min \left[\int_{\Omega} ||R(x, t+\tau) - \hat{R}(x, t+\tau)||_{L_2}^2 dx \right]$$

1.3 Primary Sources of Uncertainty and the Triad Model

To model this stochastic system, we partition the total physical uncertainty into three foundational random variables, which drive our state-space equations:

1. **The State Variable $R(x, t)$:** Rainfall intensity at location x and time t. Uncertainty arises from observational measurement error (radar reflectivity Z to rain-rate R conversion errors) and the discrete/continuous intermittency of rain.
2. **The Forcing Variable $V(x, t)$:** Motion (advection) field of rainfall patterns. This is a vector field (u, v). Uncertainty here is driven by the aperture problem in optical flow estimation, leading to trajectory divergence.

3. **The Innovation Variable $N(x,t)$:** Stochastic noise representing thermodynamic uncertainty. Even under perfect advection, convective storms spontaneously generate and dissipate. This process noise encapsulates the non-linear atmospheric chaos.

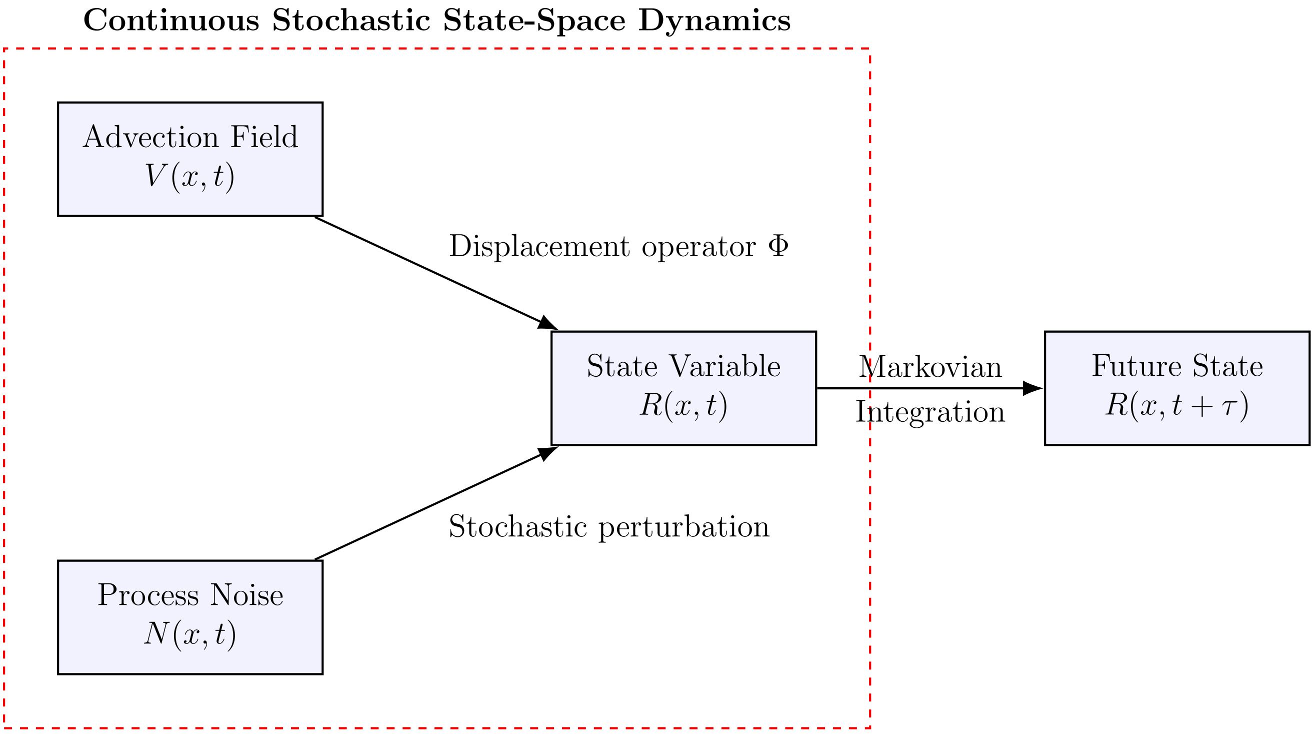

Figure 1: Conceptual Architecture of the Triad Model. The true atmospheric state $R(x,t)$ is deterministically advected by the vector field $V(x,t)$ and stochastically perturbed by the spatially correlated noise field $N(x,t)$.

Question-2: Key Random Variables and Uncertainty Modeling

Solution:

2.1 Rigorous Definition of the Random Fields

Our mathematical framework is strictly governed by the triad of continuous random fields. Let $x \in \Omega$ and $t \in \mathcal{T}$.

1. The State Variable: $R(x,t)$ **- Rainfall Intensity**

$R(x,t)$ is a continuous non-negative random field $R : \Omega \times \mathcal{T} \to^{\geq 0}$. Empirical data demonstrates that precipitation is highly positively skewed (intermittent with sudden extreme bursts). Thus, we model $R(x,t)$ as a **Log-Normal Distribution**.

To stabilize variance and allow the application of linear Gaussian mechanics, we apply a logarithmic transformation to define a latent state variable $X(x,t)$:

$$X(x,t) = 10 \log_{10}(R(x,t)) \quad \text{for } R(x,t) > R_c$$

Where R_c is a minimum observable threshold. Under this transformation, $X(x,t)$ is modeled as a Multivariate Gaussian Process:

$$X(\cdot,t) \sim \mathcal{GP}(\mu_X, \Sigma_X)$$

This transformation ensures that the extreme non-linearities of heavy rain are mapped into a linear, tractable probability space, yielding a continuous Probability Density Function (PDF):

$$f_{X_t}(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_X|}} \exp\left(-\frac{1}{2}(x-\mu_X)^T \Sigma_X^{-1} (x-\mu_X)\right)$$

2. The Forcing Variable: $V(x,t)$ **- Motion (Advection) Field**

$V(x,t) \in^2$ is a bivariate continuous random vector field representing the velocity (u,v) of the rainfall patterns. The advection field is estimated using Optical Flow equations, minimizing the brightness constancy constraint:

$$\frac{\partial R}{\partial t} + V \cdot \nabla R = 0$$

Because this partial differential equation is ill-posed (the aperture problem), $V(x,t)$ inherently contains structural uncertainty. We model $V(x,t)$ as an expected deterministic field plus a Gaussian perturbation vector:

$$V(x,t) = [V(x,t)] + \xi_V(x,t) \quad \text{where } \xi_V \sim \mathcal{N}(0, \Sigma_V)$$

3. The Innovation Variable: $N(x,t)$ **- Stochastic Noise**

$N(x,t)$ represents the sub-grid scale thermodynamic uncertainty. A crucial physical assumption is that $N(x,t)$ **cannot** be modeled as independent White Noise. If we added independent identically distributed (i.i.d.) Gaussian noise to a fluid model, the forecast would lose spatial coherence (looking like television static).

Instead, $N(x,t)$ must be a **Spatially Correlated Gaussian Random Field**:

$$N(\cdot,t) \sim \mathcal{GP}(0, \mathcal{K}(x,x'))$$

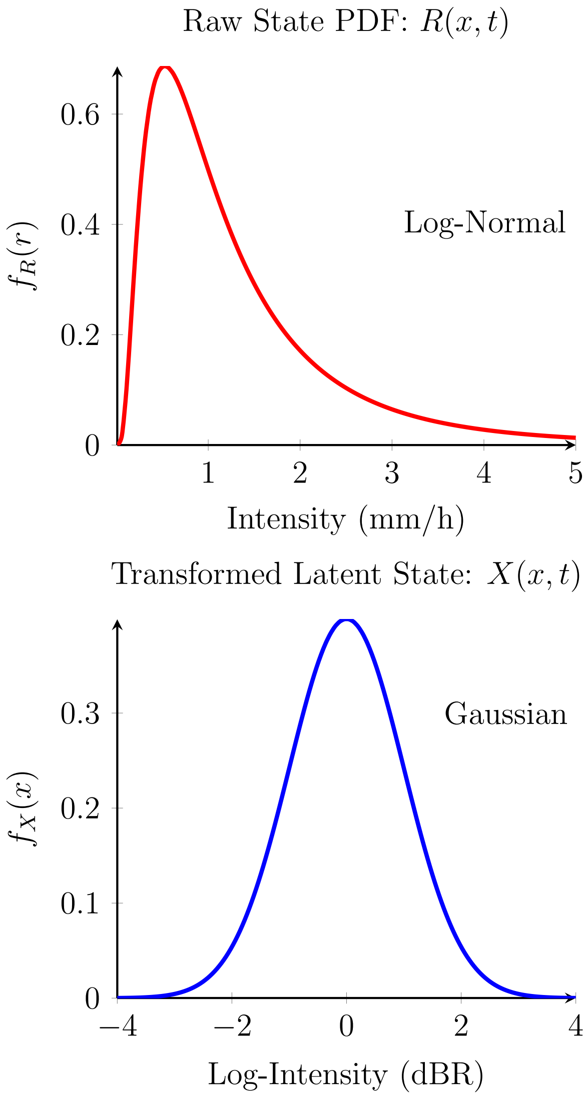

Figure 2: Probabilistic Transformation. The raw rainfall variable $R(x,t)$ is heavily skewed, breaking assumptions of linear filtering. The transformation to $X(x,t)$ establishes the strict Gaussianity required for our state-space tracking.

where $\mathcal{K}(x, x')$ is a positive-definite spatial covariance kernel. For isotropic rainfall, we assume the covariance between any two points depends only on their Euclidean distance $h = ||x - x'||_2$:

$$(N(x,t), N(x',t)) = \sigma_N^2 \rho(h)$$

Question-3: Probabilistic Reasoning and Dependencies

Solution:

3.1 Derivation of the Stochastic Partial Differential Equation (SPDE)

To model probabilistic dependencies over time, we must bridge the continuous physical PDEs with discrete stochastic processes. The physical foundation is **Lagrangian Persistence**, which states that a fluid parcel moving along a trajectory defined by $V(x,t)$ conserves its mass, barring thermodynamic sources and sinks. The material (total) derivative is:

$$\frac{DR}{Dt} = \frac{\partial R}{\partial t} + V(x,t) \cdot \nabla R = \text{Sources} - \text{Sinks}$$

In our probabilistic framework, the deterministic ”Sources/Sinks” are unobservable. Therefore, we replace them with our stochastic innovation field $N(x,t)$. Applying this to our log-transformed latent state $X(x,t)$, we yield the governing Stochastic PDE:

$$\frac{\partial X}{\partial t} + V(x,t) \cdot \nabla X = N(x,t)$$

Transitioning to discrete time steps Δt, and applying the method of characteristics along the advection trajectory, we obtain the fundamental discrete **State Transition Equation**:

$$X(x, t + \Delta t) = X(x - V(x,t)\Delta t, t) + N(x,t)$$

3.2 Temporal Dependence and the Markov Property

The discrete state equation above implies the **First-Order Markov Property**. The conditional probability measure of the future state is independent of the past history, given the current state and forcing fields:

$$(X_{t+1} \mid X_t, X_{t-1}, \ldots, X_0) = (X_{t+1} \mid X_t; V_t)$$

However, pure first-order Markov models fail to capture the long-term temporal ”memory” of mesoscale storm cells. Therefore, we embed deeper temporal dependence directly into the noise term. We model the temporal evolution of $N(x,t)$ as a stationary **Autoregressive Process of order 2, AR(2)**:

$$N(x,t) = \phi_1 N(x, t - \Delta t) + \phi_2 N(x, t - 2\Delta t) + \epsilon(x,t)$$

Where $\epsilon(x,t)$ is the fundamental, spatially correlated but temporally uncorrelated innovation, and ϕ_1, ϕ_2 are the AR parameters ensuring the system satisfies the Yule-Walker equations for stationarity.

3.3 Formal Covariance Bounds

Because spatial locations share the same advection field and spatially correlated noise, the joint distribution of any two arbitrary grid points x_i, x_j exhibits covariance:

$$(X_t(x_i), X_t(x_j)) = [X_t(x_i)X_t(x_j)] - [X_t(x_i)][X_t(x_j)] \neq 0$$

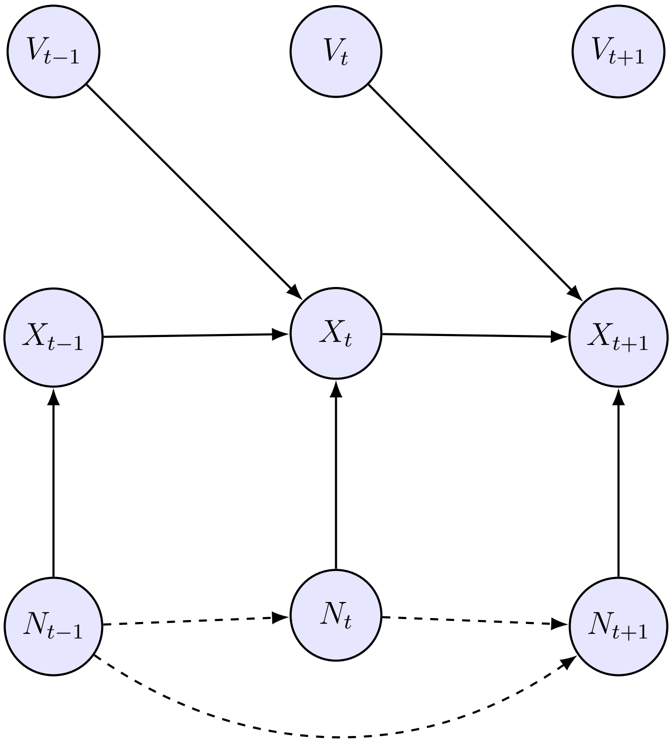

Figure 3: Dynamic Bayesian Network representing the probabilistic reasoning of the State-Space Model. Note the formal independence structures: X_{t+1} is conditionally independent of X_{t-1} given X_t, but the noise N_{t+1} depends on N_{t-1} due to the AR(2) process (dashed lines).

By Chebyshev's Inequality, for any location x, as the variance of $N(x,t)$ accumulates over time, the bounds on the forecast error strictly expand:

$$(|X_{t+\tau}(x) - [X_{t+\tau}(x)]| \geq a) \leq \frac{(X_{t+\tau}(x))}{a^2}$$

Question-4: Model–Implementation Alignment

Solution:

4.1 Computational Tractability via Spectral Theory

A severe computational bottleneck exists in generating the spatially correlated Gaussian field $N(x,t)$. For a modest 1000×1000 spatial grid, the covariance matrix Σ_N contains 10^{12} elements. Directly sampling $N \sim \mathcal{N}(0, \Sigma_N)$ requires the Cholesky Decomposition $LL^T = \Sigma_N$, which has a time complexity of $\mathcal{O}(D^3)$, rendering the implementation computationally impossible.

We align our continuous probabilistic model with discrete implementation by exploiting the **Wiener-Khinchin Theorem**. For a wide-sense stationary random field, the spatial covariance function and the Power Spectral Density (PSD) $S(k)$ form a Fourier transform pair. Therefore, we can simulate $N(x,t)$ in the frequency domain using the Fast Fourier Transform (FFT):

$$N(x,t) = \mathcal{F}^{-1}\left\{\sqrt{S(k)} \circ \mathcal{F}\{W(x,t)\}\right\}$$

where $W(x,t)$ is simple uncorrelated Gaussian White Noise, $\mathcal{F}$ represents the FFT operator, and $\circ$ is the Hadamard (element-wise) product. This aligns perfectly with the matrix-computation architecture of modern programming languages, reducing the complexity from $\mathcal{O}(D^3)$ to $\mathcal{O}(D \log D)$.

4.2 Monte Carlo Ensemble Generation and LLN

Because the advection operator $X(x - V(x,t)\Delta t, t)$ interpolates across the grid non-linearly, we cannot write closed-form analytical solutions for the posterior PDF of $X_{t+\tau}$. To solve this, we align our probability model with Monte Carlo experimentation.

We generate an ensemble of M independent realizations (stochastic trajectories). For a specific location x^* and forecast horizon τ:

$$\hat{}[X(x^*, t+\tau)] = \frac{1}{M}\sum_{m=1}^{M} X^{(m)}(x^*, t+\tau)$$

By the **Weak Law of Large Numbers (WLLN)**, since $(X^{(m)}) < \infty$, the sample mean converges in probability to the true expected value of the distribution:

$$\lim_{M\to\infty}\left(\left|\frac{1}{M}\sum_{m=1}^{M} X^{(m)} - [X]\right| \geq \epsilon\right) = 0 \quad \forall \epsilon > 0$$

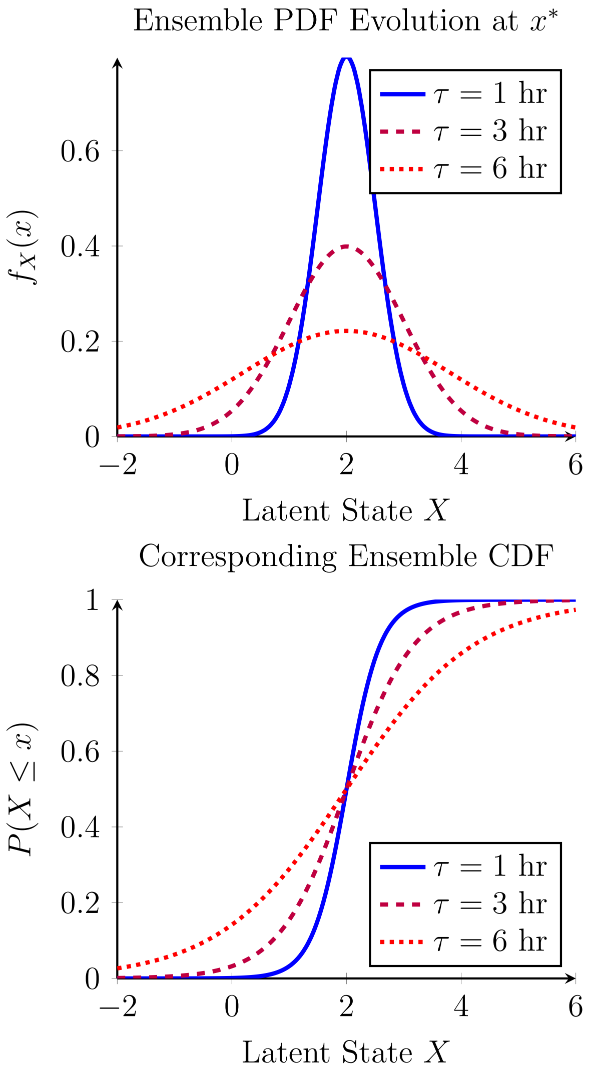

Figure 4: Simulated PDF and CDF plots generated from the Monte Carlo ensemble. As the forecast lead time (τ) increases, the continuous injection of stochastic noise $N(x,t)$ through the state equation causes the variance to grow rapidly. The PDF flattens and the CDF slope becomes shallower, representing increasing predictive uncertainty.

Question-5: Cross-Milestone Consistency and Change

Solution:

5.1 Present Well-Defined Components

At this stage, several core mathematical assumptions are fixed and serve as the immutable foundation for our state-space model:

- **The Foundational Triad:** The interaction paradigm between $R(x,t)$ (state), $V(x,t)$ (advection), and $N(x,t)$ (noise) is permanently established. The Markovian transitions will strictly rely on this interplay.
- **Log-Normal Isomorphism:** The assumption that the latent space $X(x,t) = 10\log(R)$ is Gaussian is fixed. This isomorphism allows us to utilize the superposition properties of linear combinations of normal random variables to analytically derive $(X_{t+\tau})$.
- **Spectral Decomposition:** The use of the Wiener-Khinchin theorem and FFTs to generate $N(x,t)$ is a permanent architectural choice due to its absolute necessity for computational tractability.

5.2 Aspects Expected to Evolve: Relaxing Rigid Assumptions

While the structural equations are mathematically sound, the parametrization of the specific distributions contains rigid assumptions that must be relaxed in subsequent milestones.

1. Evolution of the Vector Field: Frozen vs. Stochastic Advection

Currently, our model assumes that the advection field $V(x,t)$ estimated at $t = 0$ remains "frozen" over the forecast horizon. Mathematically, this enforces:

$$\frac{\partial V}{\partial t} = 0 \implies (V(x, t+\tau)) \approx 0$$

This is a physical fallacy. Over a 6-hour integration period, wind fields deform. In the next milestone, we must evolve $V(x,t)$ into a fully dynamic stochastic field, potentially introducing a coupled stochastic differential equation (SDE) specifically governing the velocity vectors.

2. Stationarity of Noise Covariance

Our current framework strictly assumes that the stochastic noise $N(x,t)$ is wide-sense stationary—meaning its variance σ_N^2 is identical everywhere across the spatial domain Ω:

$$(N(x_i, t)) = (N(x_j, t)) \quad \forall x_i, x_j \in \Omega$$

Atmospheric physics dictates otherwise. A severe convective thunderstorm (high R) exhibits massive thermodynamic variance, while a vast area of light stratiform rain (low R) is highly stable. In the future, we must implement **Heteroscedasticity**, defining a state-dependent covariance matrix $\Sigma_N = f(X_t)$.

Current: Frozen V

Future: Stochastic V

Figure 5: Evolution of the Forcing Variable $V(x,t)$. Currently, the advection vector (blue) is deterministic across the horizon. Future refinements will treat $V(x,t)$ probabilistically (red dashed lines indicate angular and magnitude variance in the vector field).

Question-6: Open Issues and Responsibility Attribution

Solution:

6.1 Unresolved Probabilistic Ambiguities

Despite the robust mathematical framework, applying continuous probability theory to intermittent atmospheric data presents several critical unresolved issues that break our foundational assumptions.

1. The Zero-Inflation Anomaly (Intermittency)

Our core state transformation $X(x,t) = 10\log_{10}(R(x,t))$ mathematically undefined when $R(x,t) = 0$. In reality, precipitation is a mixed discrete-continuous random variable:

$$(R = 0) \gg 0 \quad \text{and} \quad f_R(r > 0) \text{ is continuous.}$$

Our current model arbitrarily thresholds the data ($R < 0.1 \to$ Masked). This arbitrary truncation distorts the true PDF and breaks the conservation of probability mass during Monte Carlo integration. We must formalize a censored variable framework, such as a **Tobit Model**, where the true latent state X^* is allowed to be negative, but the observed state X is censored:

$$X(x,t) = \begin{cases} X^*(x,t) & \text{if } X^* > X_c \\ \text{No Rain} & \text{if } X^* \leq X_c \end{cases}$$

2. Extreme Value Theory and Fat Tails

We rely heavily on Gaussian assumptions for $X(x,t)$ and $N(x,t)$. However, meteorological data shows that extreme rainfall events (the tail of the PDF) often decay according to power-laws (e.g., Pareto or Fréchet distributions) rather than the exponential decay of a Gaussian e^{-x^2}. If our noise $N(x,t)$ has Gaussian tails, our model will systematically assign an expected probability of exactly 0.000... to 100-year flood events, violating risk-assessment requirements.

3. Topological Boundary Artifacts

The use of the Fast Fourier Transform (FFT) to simulate the correlated field $N(x,t)$ inherently imposes periodic topological boundary conditions on the spatial domain Ω. This implies Ω is mapped to a torus. A simulated storm exiting the east side of our grid mathematically re-enters on the west side. This topological artifact severely corrupts the empirical covariance calculations near the edges of the grid.

6.2 Responsibility Attribution for Next Milestone

To resolve these mathematically rigorous ambiguities, task forces for the next phase are allocated:

- **Theoretical Modeling Role:** Responsible for resolving the Zero-Inflation anomaly. This entails deriving the Maximum Likelihood Estimator (MLE) for the Tobit censoring model, preserving the continuity of the state-space equation without relying on undefined logarithmic operations.

- **Simulation & Implementation Role:** Responsible for eliminating the Boundary Condition artifact. This involves implementing domain-padding algorithms (extending the FFT domain to $\Omega' \supset \Omega$ and then truncating) to destroy the topological periodicity in $N(x,t)$.

- **Validation & Verification Role:** Responsible for auditing the Extreme Value "Fat Tail" behavior. This role must compute Continuous Ranked Probability Scores (CRPS) and generate Rank Histograms (Talagrand diagrams) comparing the ensemble CDFs against historical extremes. If the Gaussian tails are proven to be under-dispersive, they must recalibrate the innovation vector $\epsilon(x, t)$ using a stable non-Gaussian distribution.