Generative AI Modeling & Extreme Events and Heavytail distributions

Tiziano Fassina - PhD Student

T. Romary, G.V. Cardoso, S. Le Corff

What do I work on?

For my PhD Thesis, I am working on building/refining AI/Deep Learning models with the aim of better model spatial extreme events in order to better simulate and/or predict extreme scenarios.

Improved modeling, policy permitting, can help reduce the human and economic costs associated with extreme events.

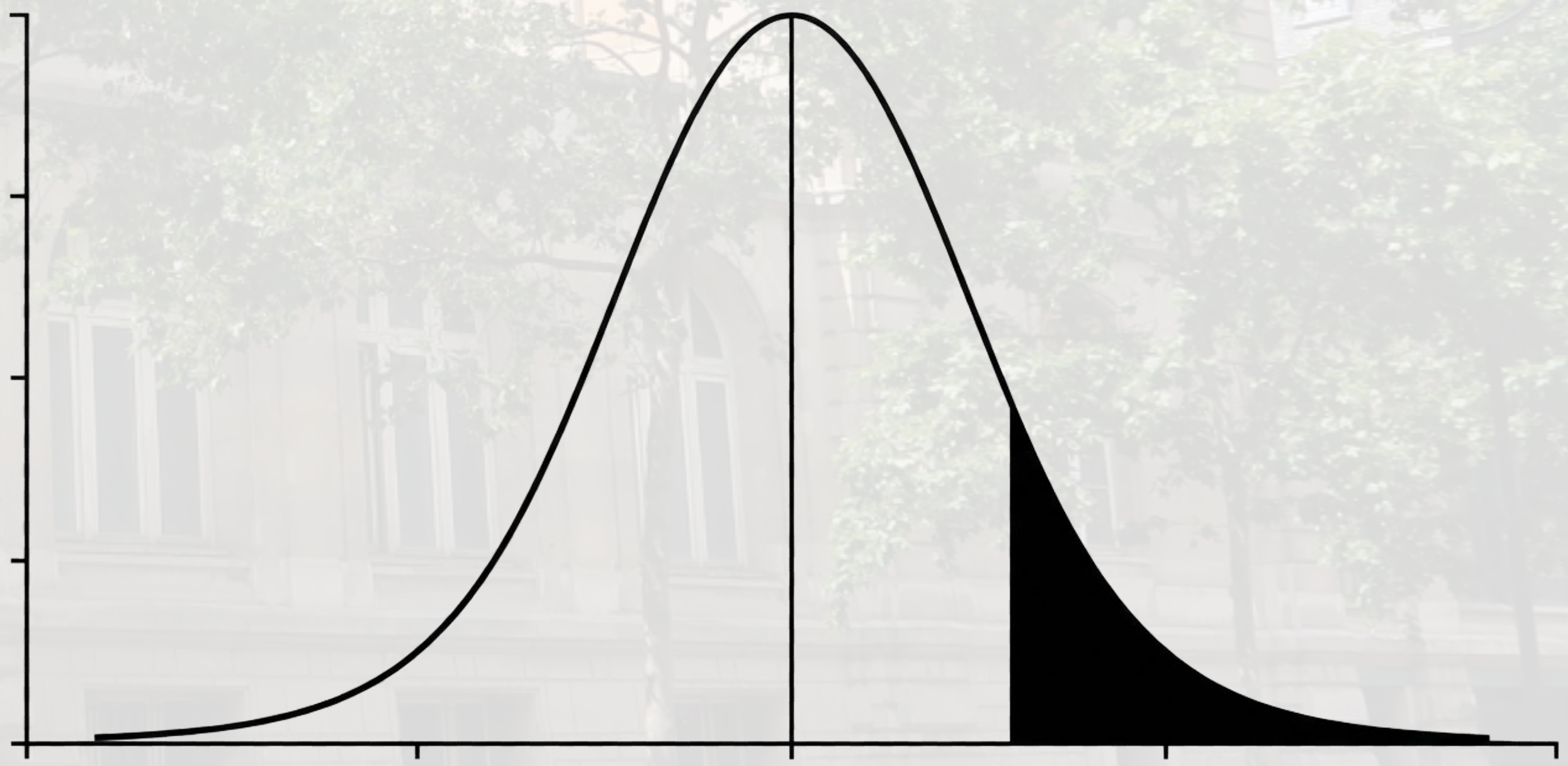
GEOLEARNING
CHAIRE /// Data Science for the Environment

What are extreme events

I work on building/refining AI/Deep Learning models with the aim of better model spatial **extreme events** with the goal of improving the simulation and prediction extreme scenarios.

Extreme events are events that have the two following features :

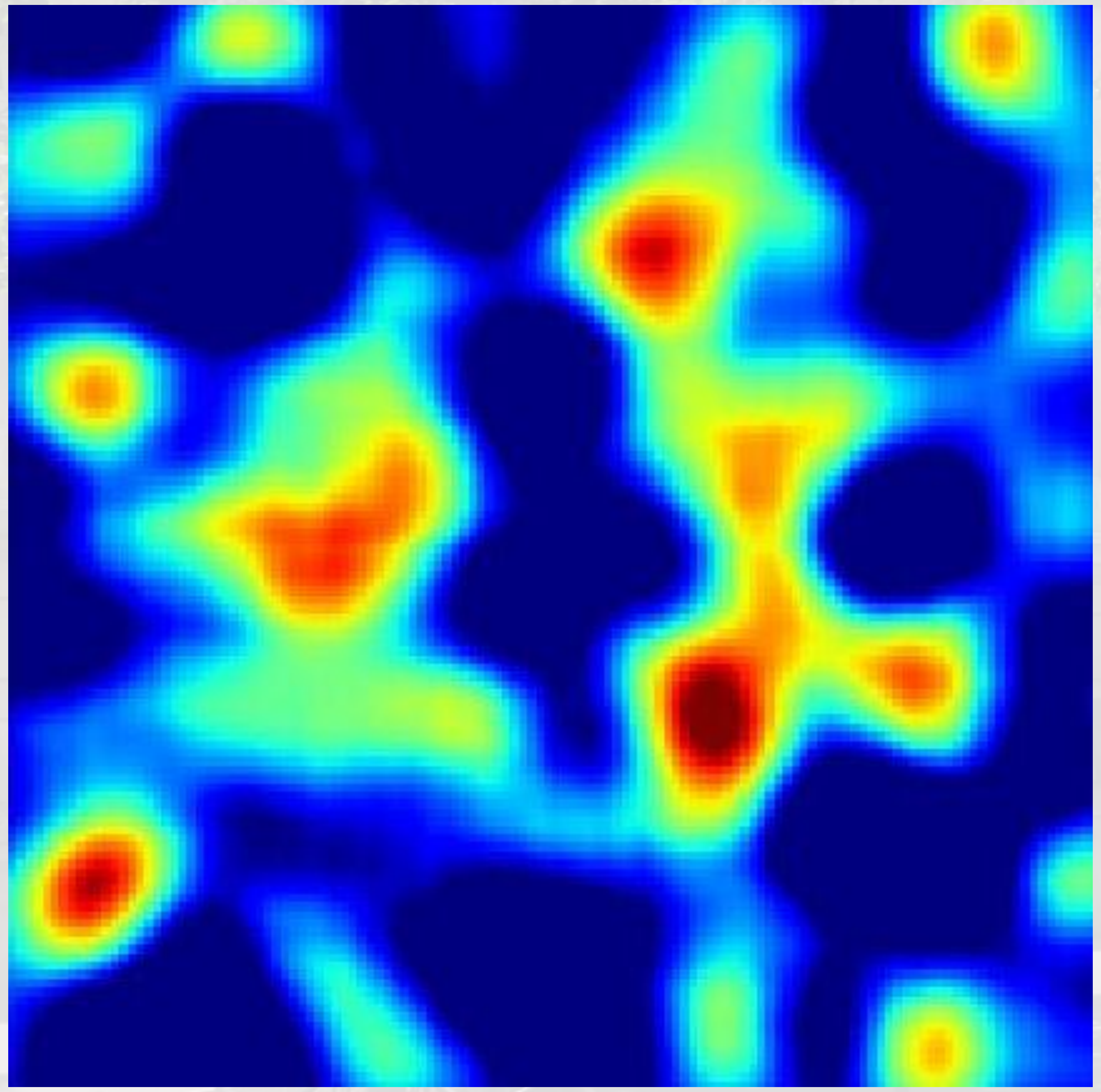
- Magnitude
- Rarity
- In probability language we talk of the tail of the distributions



GEOLEARNING
CHAIRE /// Data Science for the Environment

Spatial Data

I work on building/refining AI/Deep Learning models with the aim of better model **spatial extreme events,** with the goal of improving the simulation and prediction extreme scenarios.

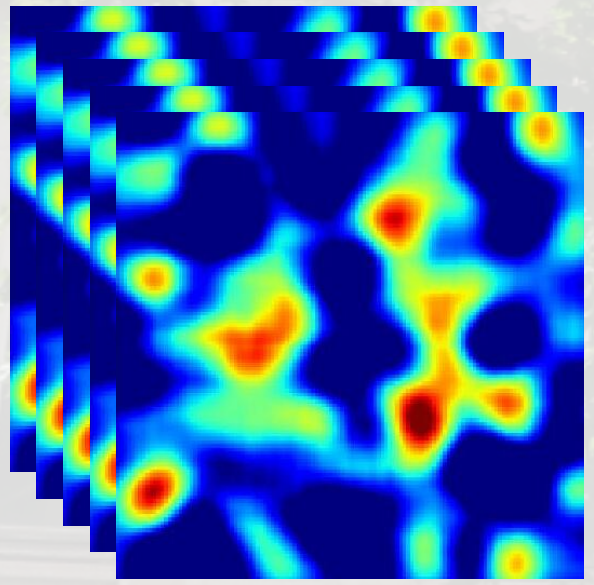


- Heuristically, image-like data
- Spatial Structure
- Rain, Temperature, Energy Demand, Epidemic diffusion, etc.
- I work and test on Rain Radar Data

GEOLEARNING
CHAIRE /// Data Science for the Environment

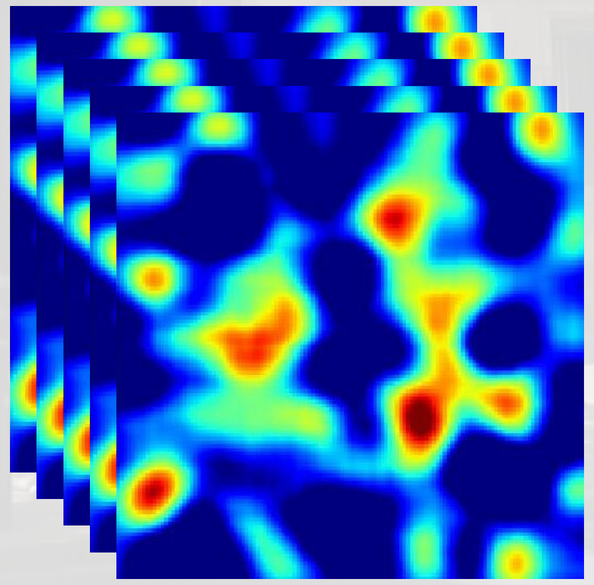
Probabilistic modeling

I work on building/refining AI/Deep Learning models with the aim of better **model spatial extreme events** with the goal of improving **the simulation and prediction** of extreme scenarios.



Data is considered as sample of a random variable X

Using data a random variable $\hat{X}$ used to generate new data



GEOLEARNING
CHAIRE /// Data Science for the Environment

Generative AI Models

I work on **building/refining AI/Deep Learning models** with the aim of better **model spatial extreme events** with the goal of improving **the simulation and prediction** of extreme scenarios.

- Diffusion Models
- Flow Matching
- Normalizing Flow
- Generative Adversarial Networks
- Variational Autoencoders

All models struggle on the tail of (heavytail) distributions and therefore in capturing extreme behaviors of distributions

GEOLEARNING
CHAIRE /// Data Science for the Environment

Generative AI Models

I work on **building/refining AI/Deep Learning models** with the aim of better **model spatial extreme events** with the goal of improving **the simulation and prediction** of extreme scenarios.

- **Diffusion Models**
- Flow Matching
- Normalizing Flow
- Generative Adversarial Networks
- Variational Autoencoders

All models struggle on the tail of (heavytail) distributions and therefore in capturing extreme behaviors of distributions

GEOLEARNING
CHAIRE /// Data Science for the Environment

Diffusion Models - the theory

I will present diffusion models in a more accessible, sufficiently rich, not exhaustive way.

I will use the Ordinary differential equation formulation.

- X_0 is a distribution with density distribution $p_0(x) : \mathbb{R}^d \to \mathbb{R}^+$
- $\sigma_t : [0,T] \to \mathbb{R}^+$ is a growing function (noise schedule)
- $p_t(x) : \mathbb{R}^d \to \mathbb{R}^+$ denotes the density of $X_t = X_0 + \sigma_t Z, \quad Z \sim \mathcal{N}(0,I_d)$
- $\dfrac{dy(t)}{dt} = \dot{\sigma}_{T-t}\sigma_{T-t}\nabla \log p_{T-t}(y(t))$

Diffusion Models - the theory

- $\frac{dy(t)}{dt} = \dot{\sigma_{T-t}}\sigma_{T-t} \nabla \log p_{T-t}(y(t)).$

- Initializing $Y(0) \sim p_T(x)$ (density of) $X_t = X + \sigma_T Z$.

- The solution of the ODE at the final point respects $Y(T) \sim p_0(x)$.

Ideally we would simulate $Y(0) \sim p_T(x)$, solve the ODE and obtain $Y(T) \sim p_0(x)$.

In practice, $p_0(x)$ is not available, therefore $p_T(x)$ and $\nabla \log p_{T-t}(x)$ are not available for any $t \in [0,T]$.

GEOLEARNING
CHAIRE /// Data Science for the Environment

Diffusion Models - approximating

- In practice we have data $\{X_i\}_i \sim p_0(x)$.
- We use data to train $s_\theta(x,t)$ such that $s_\theta(x,t) \simeq \nabla \log p_{T-t}(x)$.
- For big T we have that $X + \sigma_T Z \simeq \sigma_T Z$.
- $\dfrac{dy(t)}{dt} = \dot{\sigma}_{T-t}\sigma_{T-t}s_\theta(y(t),t)$.
- Initializing $Y(0) \sim \mathcal{N}(0,\sigma_T^2 I_d)$ or with a suitable approximation $p_\theta \simeq p_T$,
- Finding an approximated solution (by discretization) of the ODE,
- $Y(T)$ is approximately distributed as $p_0(x)$.

GEOLEARNING CHAIRE /// Data Science for the Environment

Diffusion models - a theoretical bound

$$KL(p_0, p_{Y(T)}) \leq \underbrace{KL(p_T, p_\theta)}_{\text{Initialization}} + \underbrace{\sigma_T^2 \epsilon}_{\text{Training}} + \underbrace{h\mathcal{I}(p_0)}_{\text{Discretization}}$$

$$KL(p,q) = \int_{\mathbb{R}^d} \log \frac{p(x)}{q(x)} p(x) dx$$

$$\mathcal{I}(p) = \int_{\mathbb{R}^d} \|\nabla \log p(x)\|^2 p(x) dx$$

$$\epsilon = \sup_{k=0,\ldots,N} \mathbb{E}[\|\nabla \log p_{t_k}(X_{t_k}) - s_\theta(X_{t_k}, t_k)\|^2]$$

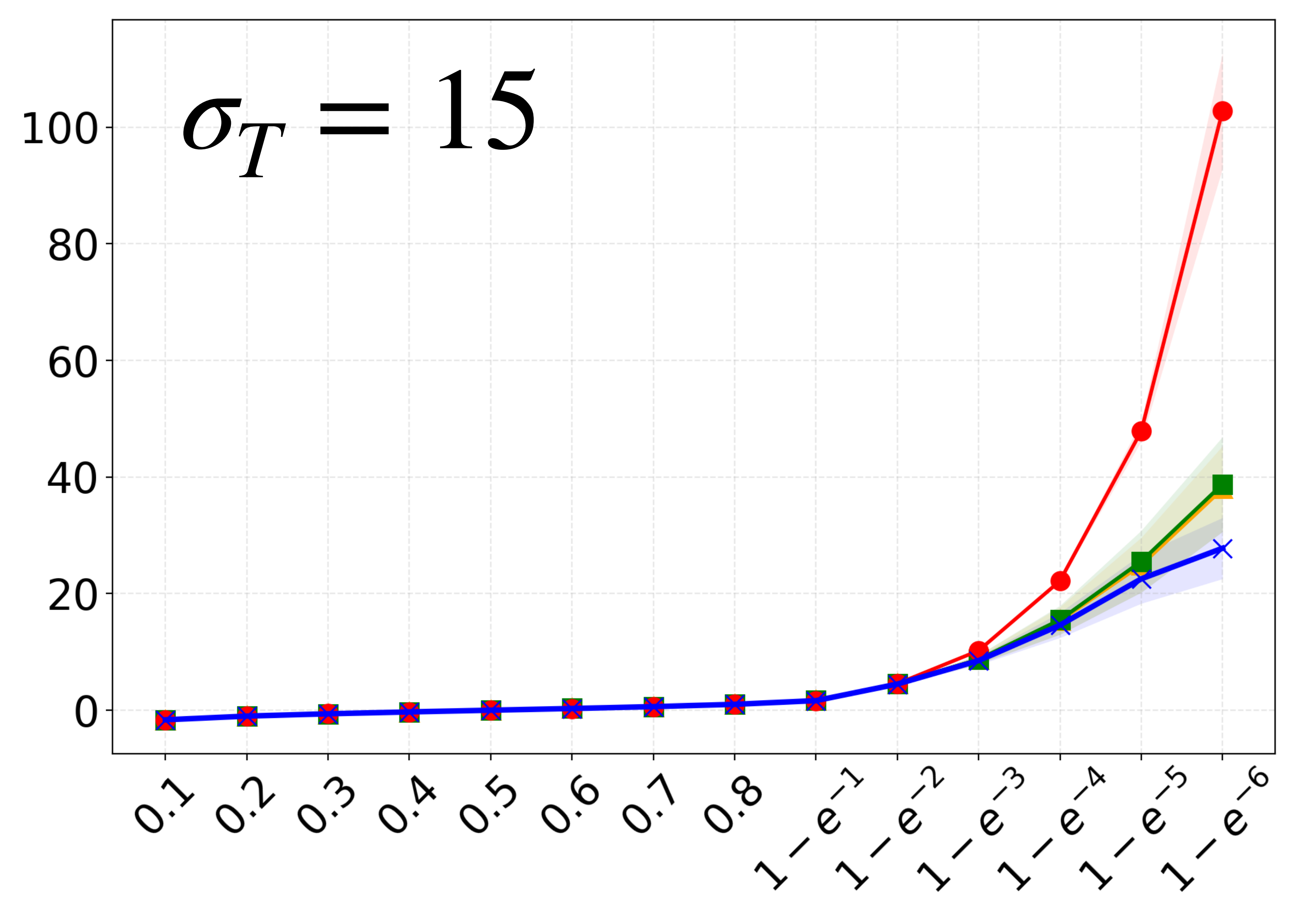
h is the discretization step of the dynamics.

* The theorem is true considering another backward scheduler. I presented diffusion and the result in this way for communication purposes. For a precise version Fassina et al, 2026, https://arxiv.org/abs/2603.00772

GEOLEARNING
CHAIRE /// Data Science for the Environment

What if we use a better initialization ?

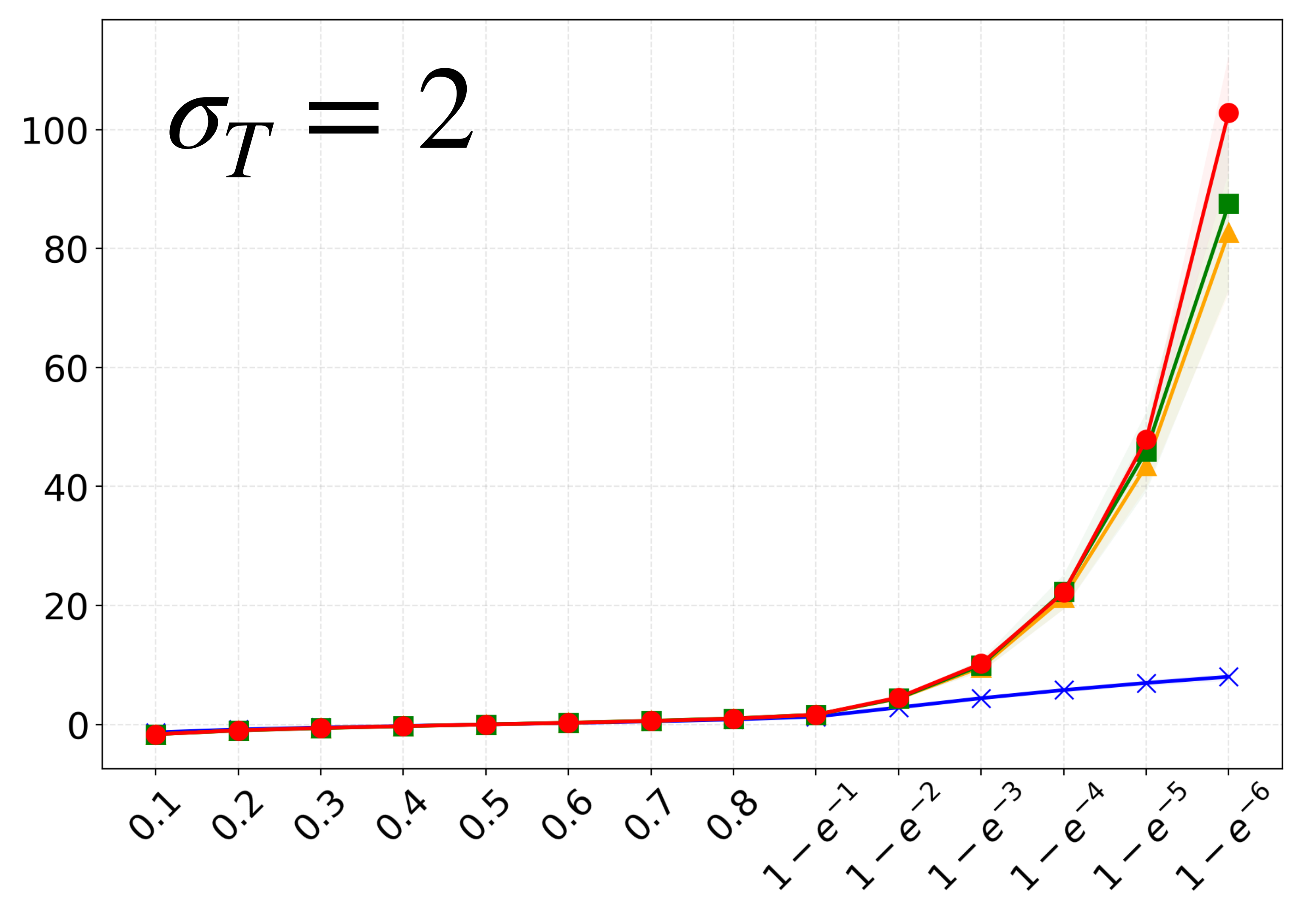
- If $p_0(x)$ is heavytail $\Longrightarrow p_T(x)$ is heavytail.
- We compare different approaches (gaussian $\mathcal{N}(0, \sigma_T^2 I_d)$, real $p_T(x)$ et learnt $p_\theta(x)$) and different time horizon σ_T in toy cases (t-Student).
- We build a quantile table using for each initialization 10^7 data. Test data is red.

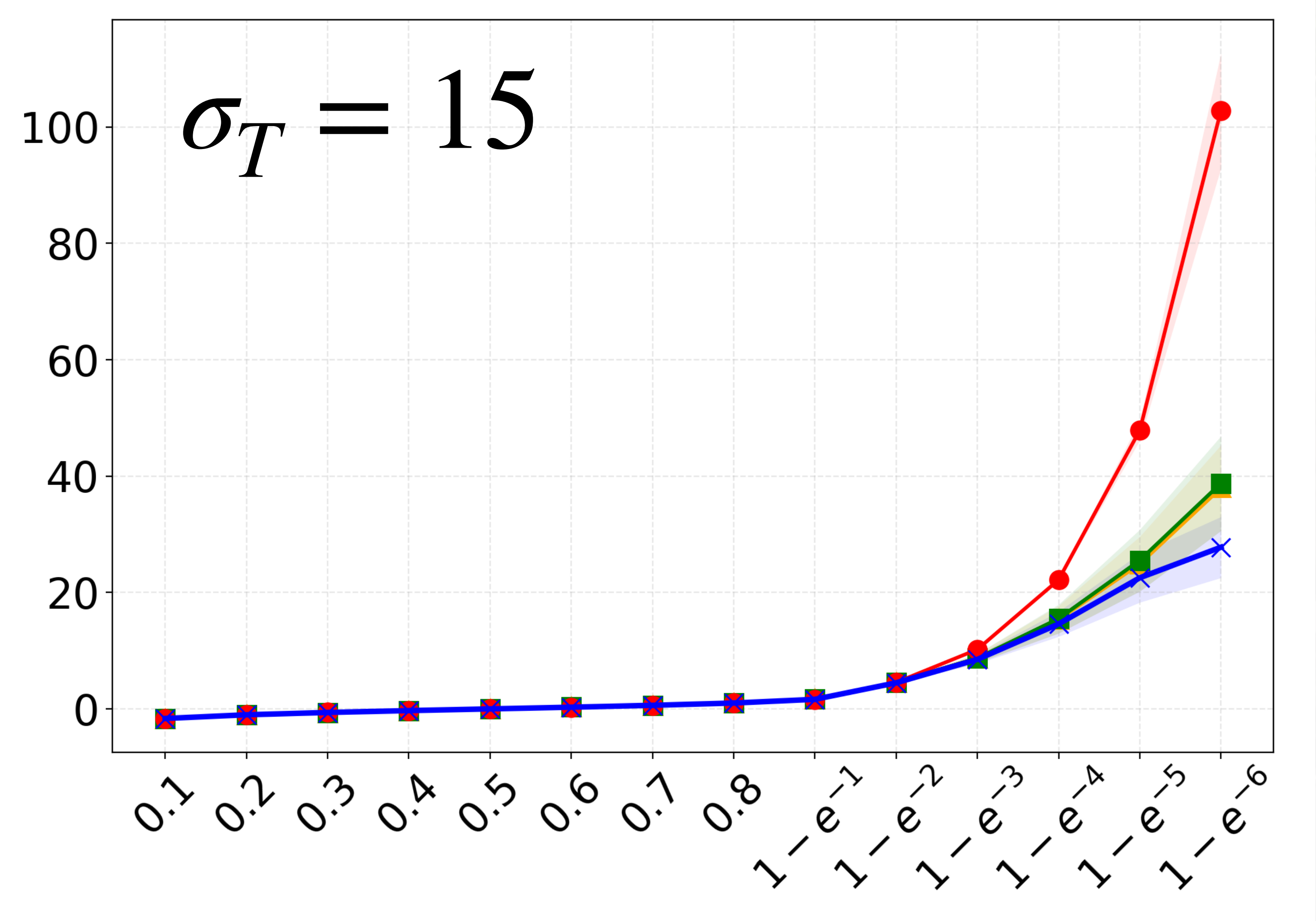


GEOLEARNING
CHAIRE /// Data Science for the Environment

What if we use a better initialization ?

- If $p_0(x)$ is heavytail $\Longrightarrow p_T(x)$ is heavytail.
- We compare different approaches (gaussian $\mathcal{N}(0, \sigma_T^2 I_d)$, real $p_T(x)$ et learnt $p_\theta(x)$) and different time horizon σ_T in toy cases (t-Student).
- We build a quantile table using for each initialization 10^7 data. Test data is red.





GEOLEARNING
CHAIRE /// Data Science for the Environment

A possible breakthrough

- **Proper initialization + a short diffusion time horizon** provides a promising approach.
- The main challenge is to design a model p_θ to capture $X_T = X_0 + \sigma_T Z$.
- Luckily, X_T is easier to model (Gaussian convolution smooths).
- Unfortunately X_T will always be heavytail.
- We explore heavy-tailed normalizing flows, good on tail - bad on details.

$$Y(0) \sim p_\theta(x) \xrightarrow{\text{Diffusion Model}} Y(T) \simeq X_0$$

Normalizing Flow

Diffusion Model

GEOLEARNING
CHAIRE /// Data Science for the Environment

Conclusion

- Modeling Heavytail Distributions is important, but standard diffusion might be limited.
- Strong tail modeling of heavy-tail normalizing flows + Detail reconstruction of diffusion models has a strong potential.
- Broad applications : climate radar downscaling, conditional simulation of extreme weather events, and forecasting severe climate scenarios.
- Also highly relevant for insurance pricing and financial market modeling, where capturing rare, extreme events is essential.

GEOLEARNING
CHAIRE /// Data Science for the Environment