Generative Cascades of Multivariate Extremes

Miguel de Carvalho

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

Joint with:

Clemente Ferrer, Tiejun Ma, Sotirios Sabanis, Ronny Vallejos *et al.*

Introducing Myself

Briefing

Credentials

- Professor at the School of Mathematics, UoE
- Chair of Statistical Data Science
- Co-Director, Edinburgh Centre for Financial Innovations
- Elected Fellow, Generative AI Laboratory

Recent Funding

Research bodies

LEVERHULME TRUST

RSE The Royal Society of Edinburgh KNOWLEDGE MADE USEFUL

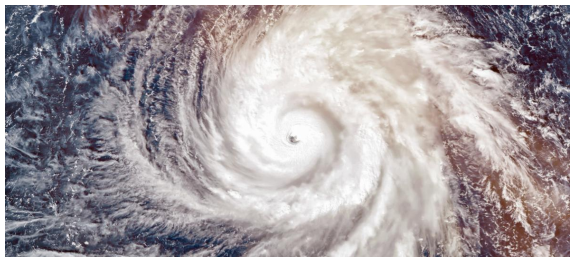
Prob_AI

Industry

aberdeen Investments

Unilever

Leading research field: Extreme Value Theory.

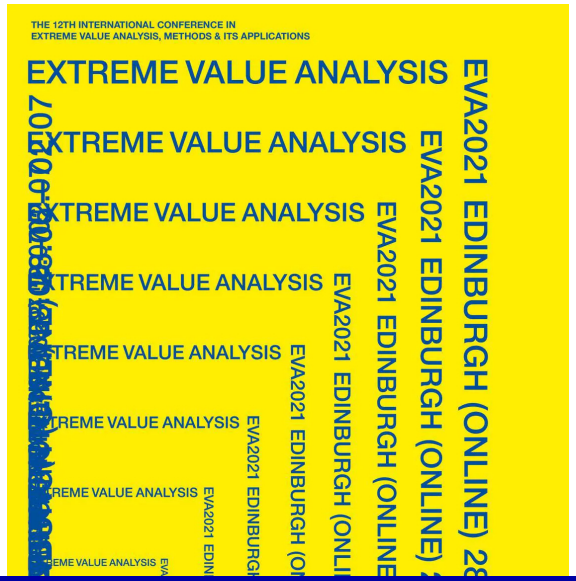

Other interests: Bayes; Interfaces between Statistics & AI.

Introducing Myself

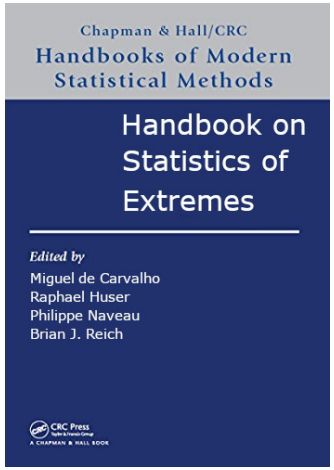
GAME 2025

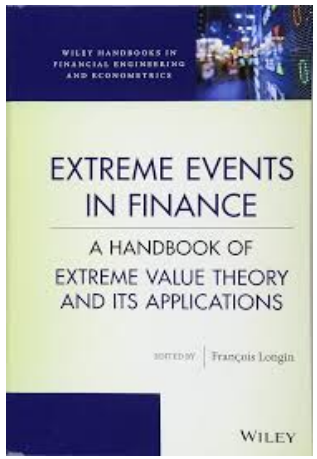
Extremes, Special Issue (2026)

EVA 2021



Handbooks





Introduction and Motivation

Vision

- Generative AI has been entering a new era. Many opportunities beyond text, images, video, etc.
- I believe it is important for us to position ourselves at the frontier of this shift. This is the natural next step.

Structure of today's talk

- Cascading Extremes.
- Generative Transformer-Based Approaches for Cascading Extremes.

Introduction and Motivation

Vision

- Generative AI has been entering a new era. Many opportunities beyond text, images, video, etc.
- I believe it is important for us to position ourselves at the frontier of this shift. This is the natural next step.

Structure of today's talk

- Cascading Extremes.
- Generative Transformer-Based Approaches for Cascading Extremes.

Vision

A cutting-edge risk engine for generative cascades of extremes.

Extremes manuscript No.
(will be inserted by the editor)

A Kolmogorov–Arnold Neural Model for Cascading Extremes

Miguel de Carvalho[1,2] · **Clemente Ferrer**[3] · **Ronny Vallejos**[3]

Received: date / Accepted: date

Abstract This paper addresses the growing concern of cascading extreme events, such as an extreme earthquake followed by a tsunami, by presenting a novel method for risk assessment focused on these domino effects. The proposed approach develops an extreme value theory framework within a Kolmogorov–Arnold network (KAN) to estimate the probability of one extreme event triggering another, conditionally on a feature vector. An extra layer is added to the KAN architecture to ensure that the parameter of interest lies within the unit interval, and we refer to the resulting neural model as KANE (KAN with Natural Enforcement). The proposed method is backed by exhaustive numerical studies and further illustrated with real-world applications to seismology and climatology.

Keywords Bernoulli process, Chained extreme events, KAN, Kolmogorov superposition theorem, Neural network, Multivariate extremes, Regression models for extremes

Part I

Neural Statistical Modeling of Cascading Extremes

Introduction and Motivation

Compound, Cascading, and Complex Extreme Events

- While it is widely recognized by practitioners that extreme events tend to occur in complex sequential forms (Cutter2018; Raymond et al.2020), statistical modelling of such context from an EVT viewpoint is still underdeveloped.
- Multivariate EVT, though a natural starting point, falls short by:
  - disregarding the triggering role of certain events;
  - overlooking the order and sequential nature of extreme event cascades;
  - lacking the ability to model feedback loops between events.

Introduction and Motivation

The POC Surface

- Inspired by the multivariate EVT framework, in this talk I introduce a novel concept, the POC (Probability of Cascade) surface, to assess the probability of domino effects between extreme events conditionally on a covariate or feature vector $\mathbf{x} \in \mathbb{R}^p$.
- The proposed POC-based approach is fully general in the sense that the focus can be placed beyond the case where follow-up event is binary.
- In particular, we extend the framework to a multi-class setting, allowing for different types of follow-up extreme events.
- The case where the follow-up event is continuous includes as a particular case the conditional coefficient of extremal dependence introduced by Lee et al.2024).

Background

- To learn about the POC surface from the data, we develop a neural model grounded on Kolmogorov's superposition theorem.

Background

- To learn about the POC surface from the data, we develop a neural model grounded on Kolmogorov's superposition theorem.
- Superpositions are functions of functions.

Background

- To learn about the POC surface from the data, we develop a neural model grounded on Kolmogorov's superposition theorem.
- Superpositions are functions of functions.

Example (Superposition of univariate and bivariate functions)

$$f(x_1, x_2, x_3) = g\big(a(\alpha(x_1), \beta(x_2, x_3)), b(x_1, x_2)\big).$$

Theorem (Kolmogorov's superposition theorem)

Let $f : [0,1]^d \to \mathbb{R}$ *be a continuous function.*

Background

- To learn about the POC surface from the data, we develop a neural model grounded on Kolmogorov's superposition theorem.
- Superpositions are functions of functions.

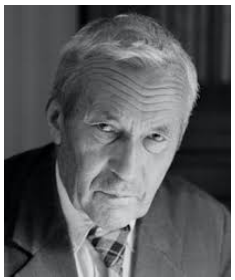
Example (Superposition of univariate and bivariate functions)

$$f(x_1, x_2, x_3) = g\big(a(\alpha(x_1), \beta(x_2, x_3)), b(x_1, x_2)\big).$$

Theorem (Kolmogorov's superposition theorem)

Let $f : [0,1]^d \to \mathbb{R}$ be a continuous function. Then,

$$f(\mathbf{x}) = \sum_{i=1}^{2d+1} \Phi_i^{(2)} \left(\sum_{j=1}^{d} \Phi_{i,j}^{(1)}(x_j) \right), \qquad \mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}},$$



A. Kolmogorov

for some continuous one-dimensional functions $\Phi_{i,j}^{(1)}$ and $\Phi_j^{(2)}$.

Background

- From an AI perspective, the theorem reveals a two-layer neural network architecture recently popularized by Liu et al.2024) following their extension to deeper settings.
- While multi-layer perceptrons are inspired by the universal approximation theorem (e.g., Berlyand and Jabin2023), Kolmogorov–Arnold Networks (KAN) are a novel and fast-evolving addition to the AI toolbox, and are rooted on Kolmogorov's superposition theorem.
- A particularly impressive aspect of KAN is that they are based on the principle any multivariate continuous function can be expressed exactly using only $2d + 1$ outer functions and d inner functions.
- In addition to the many developments following Liu et al., it should be noted that other neural approaches based on this theorem had already appeared in the literature (Lin and Unbehauen1993; Sprecher and Draghici2002; Montanelli and Yang2020; Fakhoury et al.2022).

Cascading Extremes

Modeling Chained Extreme Events

- Let $I = \{I_u : u \in \mathbb{R}\}$ be a Bernoulli process and $Y \sim F_Y$ be a continuous rv.
- We start by introducing the following functional, referred to as alpha, which plays a central role in our developments:

$$\alpha \equiv \alpha_I = \lim_{u \to y^*} P(I_u = 1 \mid Y > u).$$

- Notation: $y^* = \sup\{y : F_Y(y) < 1\}$ is the right endpoint of F_Y.
- Loosely, α is the probability of a follow-up event (like a tsunami $I_u = 1$), given a trigger event (such as an earthquake exceeding magnitude u).
- The nature of the Bernoulli process I defining the follow-up event opens up a variety of modeling possibilities as illustrated below.

Examples of Alpha

Example (Tail dependence coefficient)

If $I_u = I(Z > u)$, where Y and Z have common distribution, then

$$\alpha^{\text{TDC}} \equiv \alpha_I = \lim_{u \to y^*} P(Z > u \mid Y > u).$$

Thus, α includes the well-known tail dependence coefficient as a special case when the follow-up event involves Z being extreme and observable. ■

Example (Extremal probabilistic index)

If $I_u = I(Z > Y)$, then

$$\alpha^{\text{PI}} \equiv \alpha_I = \lim_{u \to y^*} P(Z > Y \mid Y > u),$$

which can be regarded as extremal version of the probabilistic index (Thas et al.2012). ■

POC Surface

Setup and Definition

- Our setup keeps in mind that for some applications the variable Z in the above examples might be latent, but it assumes that the Bernoulli process I is always observable.
- In practice it is desirable to assess how the α functional may be impacted by a covariate or feature.

Definition (POC Surface)

Let $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathcal{X} \subseteq \mathbb{R}^d$. The probability of cascade surface is defined as

$$\mathrm{POC} = \{(\mathbf{x}, \alpha_I(\mathbf{x})) : \mathbf{x} \in \mathcal{X}\}, \quad \alpha_I(\mathbf{x}) = \lim_{u \to y^*} P(I_{u,\mathbf{x}} = 1 \mid Y_{\mathbf{x}} > u),$$

where $I = \{I_{u,\mathbf{x}} : (u, \mathbf{x}) \in \mathbb{R} \times \mathcal{X}\}$ is a random field with Bernoulli marginal distributions and $\{Y_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ is a random field.

POC Surface

A Kolmogorov–Arnold Approach for Learning from Data

Starting Point (KAN):

$$\alpha_l(\mathbf{x}) = \sum_{i=1}^{2d+1} \Phi_i^{(2)} \left(\sum_{j=1}^{d} \Phi_{i,j}^{(1)}(x_j) \right).$$

Or in function matrix notation

$$\alpha_l(\mathbf{x}) = (\mathbf{\Phi}^{(2)} \circ \mathbf{\Phi}^{(1)})\,\mathbf{x},$$

where

$$\mathbf{\Phi}^{(1)} = \begin{pmatrix} \Phi_{1,1}^{(1)} & \cdots & \Phi_{1,d}^{(1)} \\ \vdots & \ddots & \vdots \\ \Phi_{2d+1,1}^{(1)} & \cdots & \Phi_{2d+1,d}^{(1)} \end{pmatrix}, \qquad \mathbf{\Phi}^{(2)} = \begin{pmatrix} \Phi_1^{(2)} \\ \vdots \\ \Phi_{2d+1}^{(2)} \end{pmatrix}^{\mathrm{T}}.$$

Issue: $\alpha_l(\mathbf{x})$ may not be in $[0, 1]$!

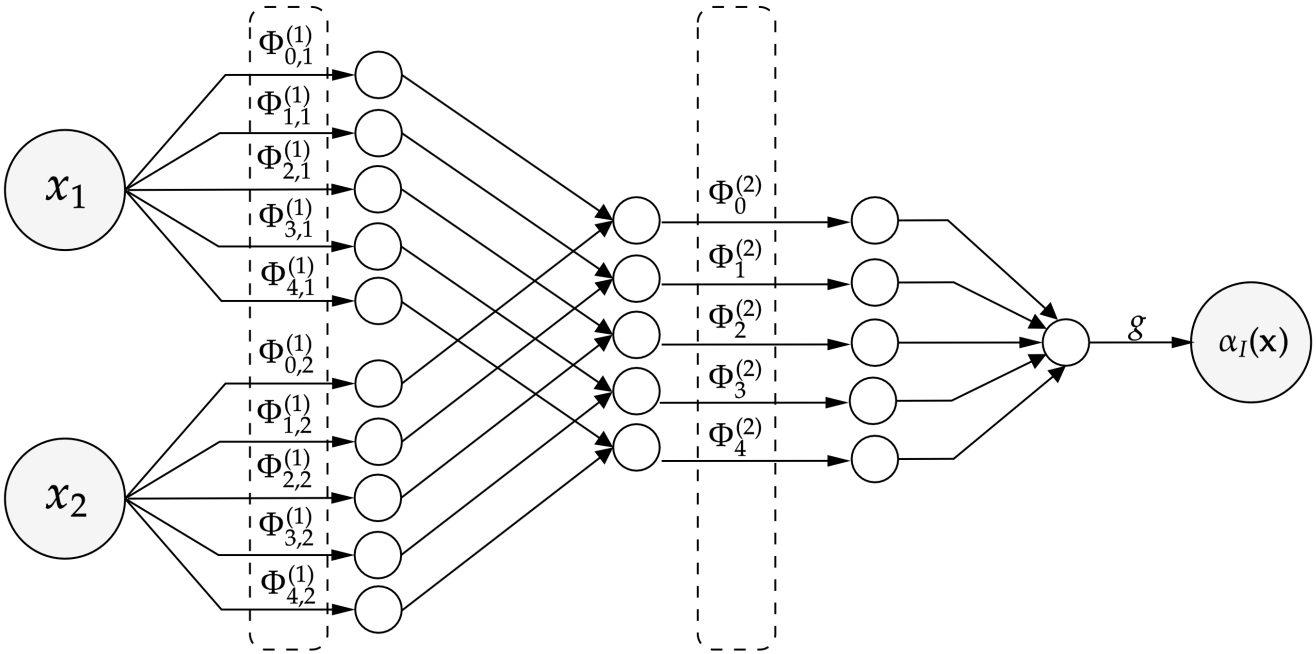
POC Surface

KANE: **KAN** with Natural **E**nforcement

Refined Version (KANE): Let $g : \mathbb{R} \to [0, 1]$, and set

$$\alpha_l(\mathbf{x}) = g\left(\sum_{i=1}^{2d+1} \Phi_i^{(2)}\left(\sum_{j=1}^{d} \Phi_{i,j}^{(1)}(x_j)\right)\right).$$

Or in function matrix notation $\alpha_l(\mathbf{x}) = g(\mathbf{\Phi}^{(2)} \circ \mathbf{\Phi}^{(1)})\,\mathbf{x}$. **Now**: $\alpha_l(\mathbf{x}) \in [0, 1]$.



Deep POC Surface

A Deep Version of the Model *a la* Liu et al.

Deep Version (L Layer Model): Let $g : \mathbb{R} \to [0, 1]$, and set

$$\alpha_l(\mathbf{x}) = g(\mathbf{\Phi}^{(L-1)} \circ \cdots \circ \mathbf{\Phi}^{(1)}),$$

where

$$\mathbf{\Phi}^{(l)} = \begin{pmatrix} \Phi_{1,1}^{(l)} & \cdots & \Phi_{1,n_l}^{(l)} \\ \vdots & \ddots & \vdots \\ \Phi_{n_l+1,1}^{(l)} & \cdots & \Phi_{n_{l+1},n_l}^{(l)} \end{pmatrix},$$

where n_l is the number of nodes in the lth layer.

Deep POC Surface

A Kolmogorov–Arnold Approach for Learning from Data

- Consider $m+1$ equally-spaced knots, $t_0 < \cdots < t_m$. We model the inner and outer functions as a linear combination of B-spline basis functions, that is,

$$\Phi_{i,j}^{(l)}(x) = \sum_{k=1}^{K} \beta_{i,j,k}^{(l)} B_k^p(x),$$

for $i = 1, \ldots, d$, where $B_k^p(x)$ is a B-spline basis function of degree p evaluated at x and $K = p + m$.

- The parameter of interest is given by the following collection of matrices

$$\boldsymbol{\beta}_k^{(l)} = \begin{pmatrix} \beta_{n_l,1,k}^{(l)} & \cdots & \beta_{n_l,d,k}^{(l)} \\ \vdots & \ddots & \vdots \\ \beta_{n_{l+1},1,k}^{(l)} & \cdots & \beta_{n_{l+1},d,k}^{(l)} \end{pmatrix},$$

where $k = 1, \ldots, K$ and $l = 1, \ldots, L$.

Consequences & Extensions

Multi-Trigger Systems

- In practice, multiple competing incidents can contribute to trigger in the follow-up event.
- To address this we define a multi-trigger system where we consider $\mathcal{Y}_1, \ldots, \mathcal{Y}_K$ as a sequence of identically distributed random fields with

$$\mathcal{Y}_k = \{Y_{k,\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}, \qquad k = 1, \ldots, K.$$

- Our framework extends to the framework of K trigger events by considering

$$\alpha_I(\mathbf{x}) = \lim_{u \to y^*} P(I_{u,\mathbf{x}} = 1 \mid Y_{1,\mathbf{x}} > u \vee \cdots \vee Y_{K,\mathbf{x}} > u).$$

- This formula simplifies to the original single-trigger case by defining $Y_{\mathbf{x}} = \min\{Y_{1,\mathbf{x}}, \ldots, Y_{K,\mathbf{x}}\}$, and hence the theory and methods discussed earlier readily apply to this context as well.

Consequences & Extensions

Categorical, Ordinal, and Continuous Follow-up Events

- In real-world applications, the follow-up extreme event may come in different flavors or categories.

Example

For example, $j = 0$ may represent *no tornado*, $j = 1$, *supercell tornado*, and $j = 2$ a *non-supercell tornado*.

- The proposed cascading probability surfaces naturally extend to this context as follows

$$\alpha_I(\mathbf{x})^{(j)} = \lim_{u \to y^*} P(I_{u,\mathbf{x}} = j \mid Y_{\mathbf{x}} > u),$$

where $j = 0, \ldots, J$.

Theoretical Properties

The Kolmogorov Superposition Operator

Highlight

As shown below small perturbations in the inner and outer functions induce only small perturbations in the resulting function.

- Throughout, $\|f\|_\infty \equiv \|f\|_\infty^A := \sup_{x\in A} |f(x)|$, and $C(A)$ and $C_{\text{Lip}}(A)$ denote the spaces of continuous and Lipschitz continuous functions on $A \subseteq \mathbb{R}$.
- Let $I = \{1, \ldots, 2d+1\}$ and $J = \{1, \ldots, d\}$.
- Finally, we equip $C([0,1])^{(2d+1)d} \times C_{\text{Lip}}(\mathbb{R})^{2d+1}$ with the max norm

$$\|\mathbf{\Phi}\| = \|(\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)})\| = \max(m_1, m_2),$$

where

$$m_1 = \max_{(i,j)\in I\times J} \|\Phi_{i,j}^{(1)}\|_\infty, \qquad m_2 = \max_{i\in I} \|\Phi_i^{(2)}\|_\infty.$$

Theoretical Properties

Continuity of Kolmogorov Superposition Operator

Theorem

Consider the operator

$$\mathscr{K} : C([0,1])^{d(2d+1)} \times C_{Lip}(\mathbb{R})^{2d+1} \rightarrow C([0,1]^d),$$

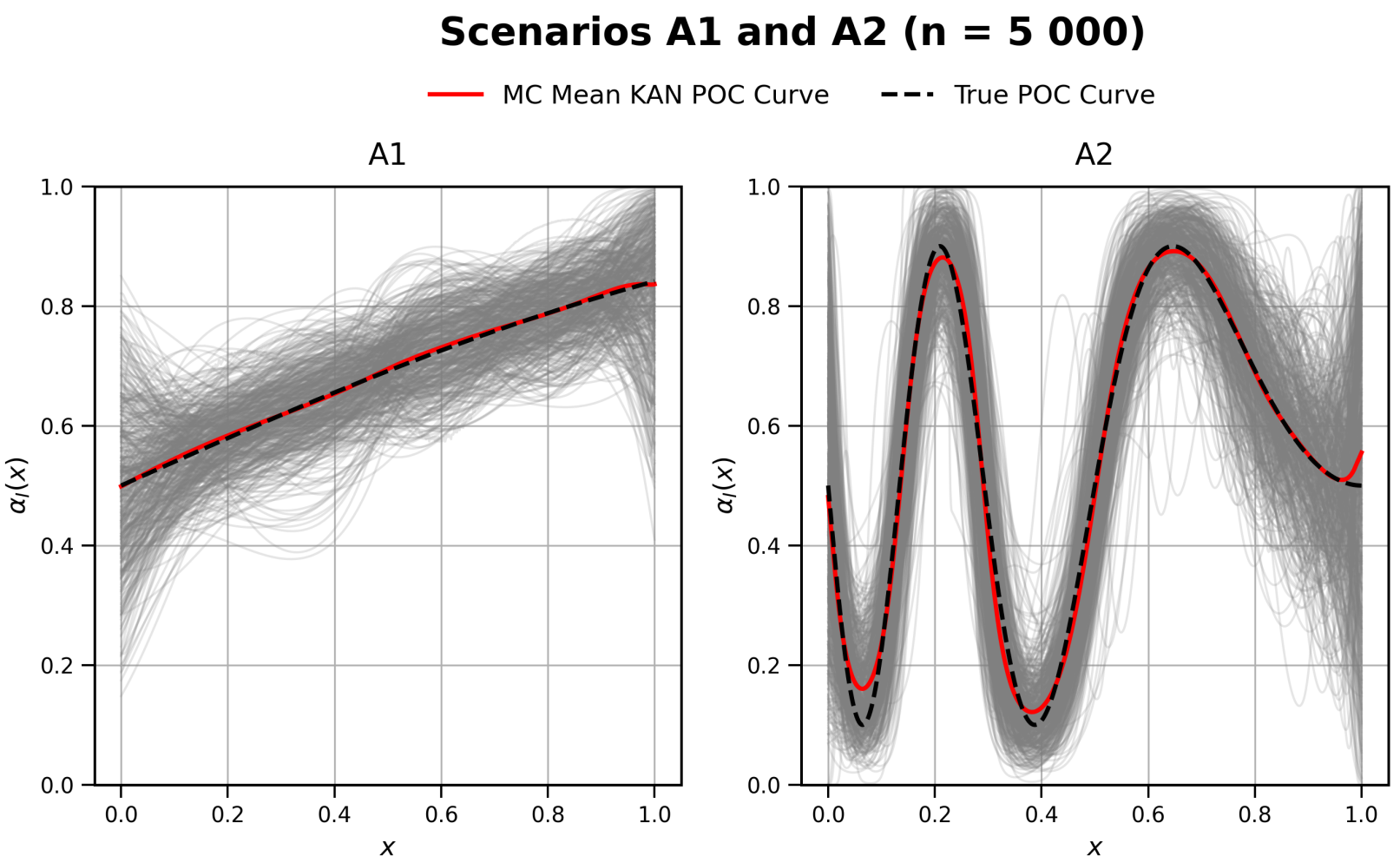
defined by $\mathscr{K}(\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)})(x_1, \ldots, x_d) = \sum_{i=1}^{2d+1} \Phi_i^{(2)}(\sum_{j=1}^{d} \Phi_{i,j}^{(1)}(x_j))$, *with*

$$\mathbf{\Phi}^{(1)} = \left(\Phi_{i,j}^{(1)}\right)_{(i,j)\in I\times J} \in C([0,1])^{d(2d+1)}, \qquad \mathbf{\Phi}^{(2)} = \left(\Phi_i^{(2)}\right)_{i\in I} \in C_{Lip}(\mathbb{R})^{2d+1},$$

and $I = \{1, \ldots, 2d+1\}$ *and* $J = \{1, \ldots, d\}$. *Then,* $\mathscr{K}$ *is a continuous operator.*

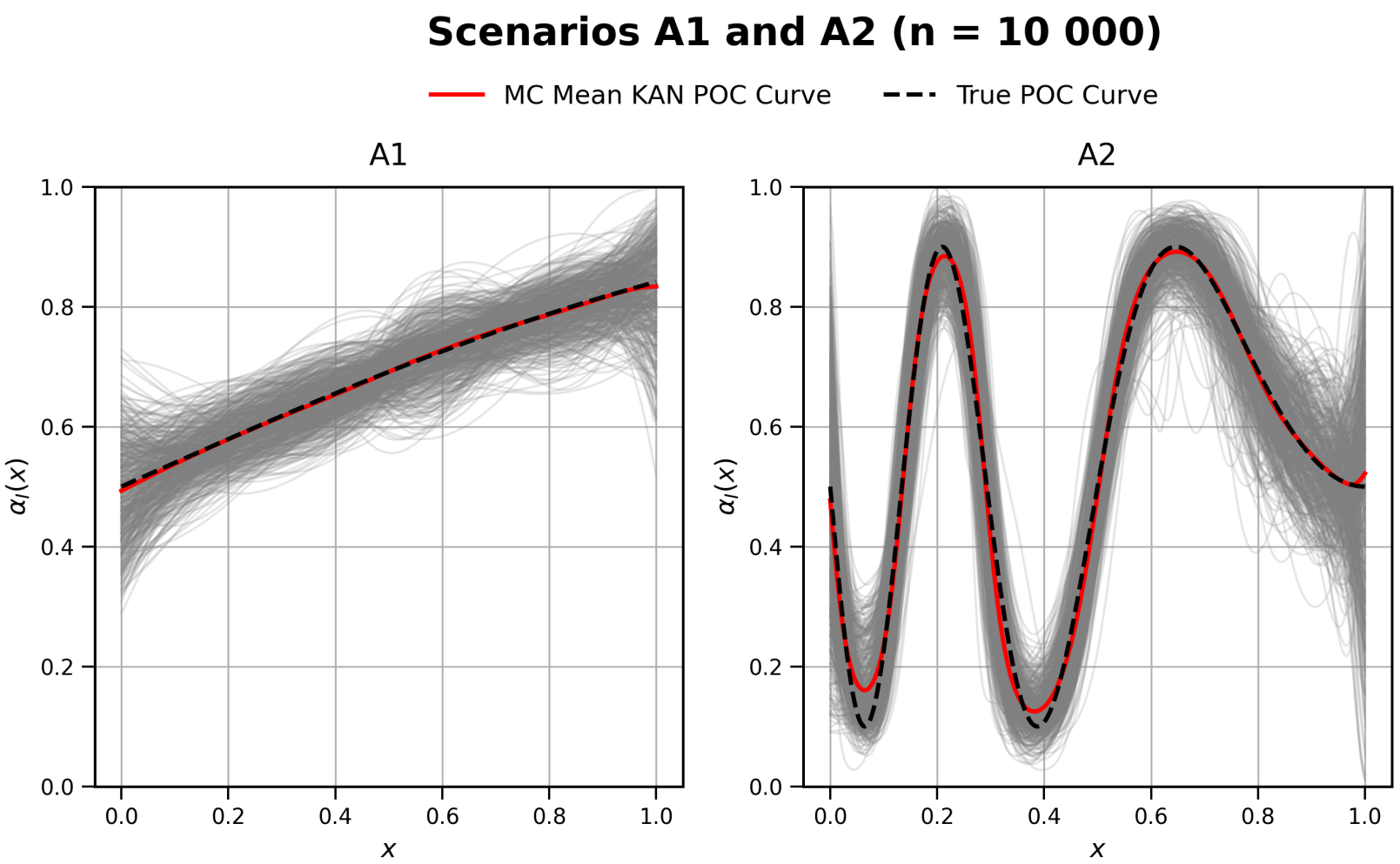
Monte Carlo Simulation Study

Scenarios A



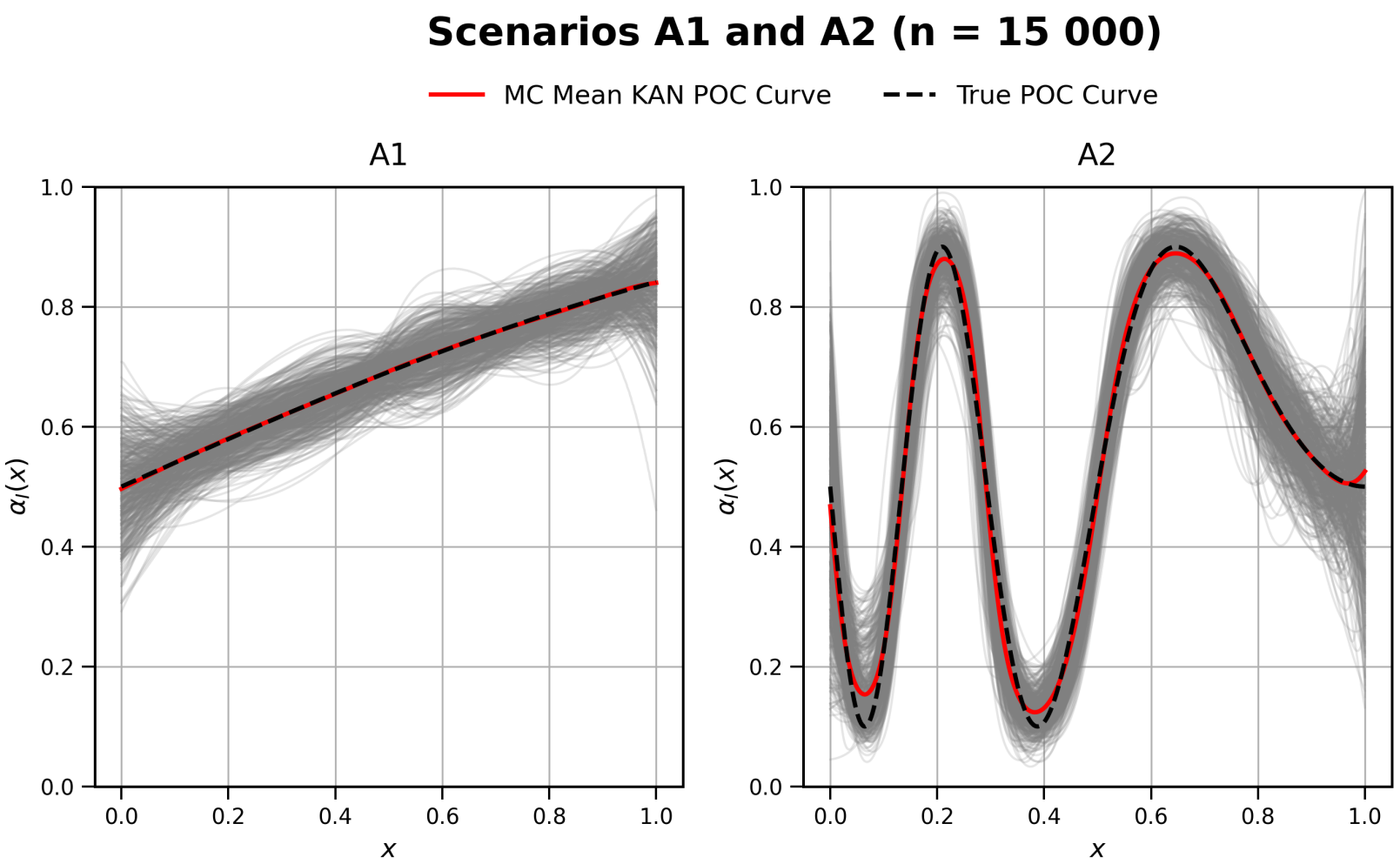
Monte Carlo Simulation Study

Scenarios A



Monte Carlo Simulation Study

Scenarios A



Real Data Illustration

Applied Rationale and Data Description

- Coastal regions are highly vulnerable to tsunamis, with their impacts often compounded by extreme earthquakes that act as primary triggers.
- We now apply the proposed method to quantify the probability of cascade for tsunami occurrence triggered by extreme earthquakes.
- Our analysis uses data from the NCEI/WDS Global Significant Earthquake Database, provided by the NOAA National Centers for Environmental Information.
- The dataset contains over 5 700 significant earthquakes from 2150 B.C. to present, defined by criteria such as fatalities, damages over $1 million, Modified Mercalli Intensity (MMI) X or greater, or the earthquake generated a tsunami.
- Each observation includes event date, location, depth, magnitude, MMI, and socio-economic impacts (casualties, injuries, property damage), with references and notes on related events like tsunamis and eruptions.

Real Data Illustration

Implementation

- We threshold the data at their 95% threshold for fitting the POC surface and consider the features,

$$(\text{latitude}, \text{longitude}, \text{depth}).$$

- We transform longitude and latitude coordinates to the unit square for modeling purposes.
- Finally, we fit a KAN model with three layers, using a sigmoid activation function at the output.

Real Data Illustration

Visualization

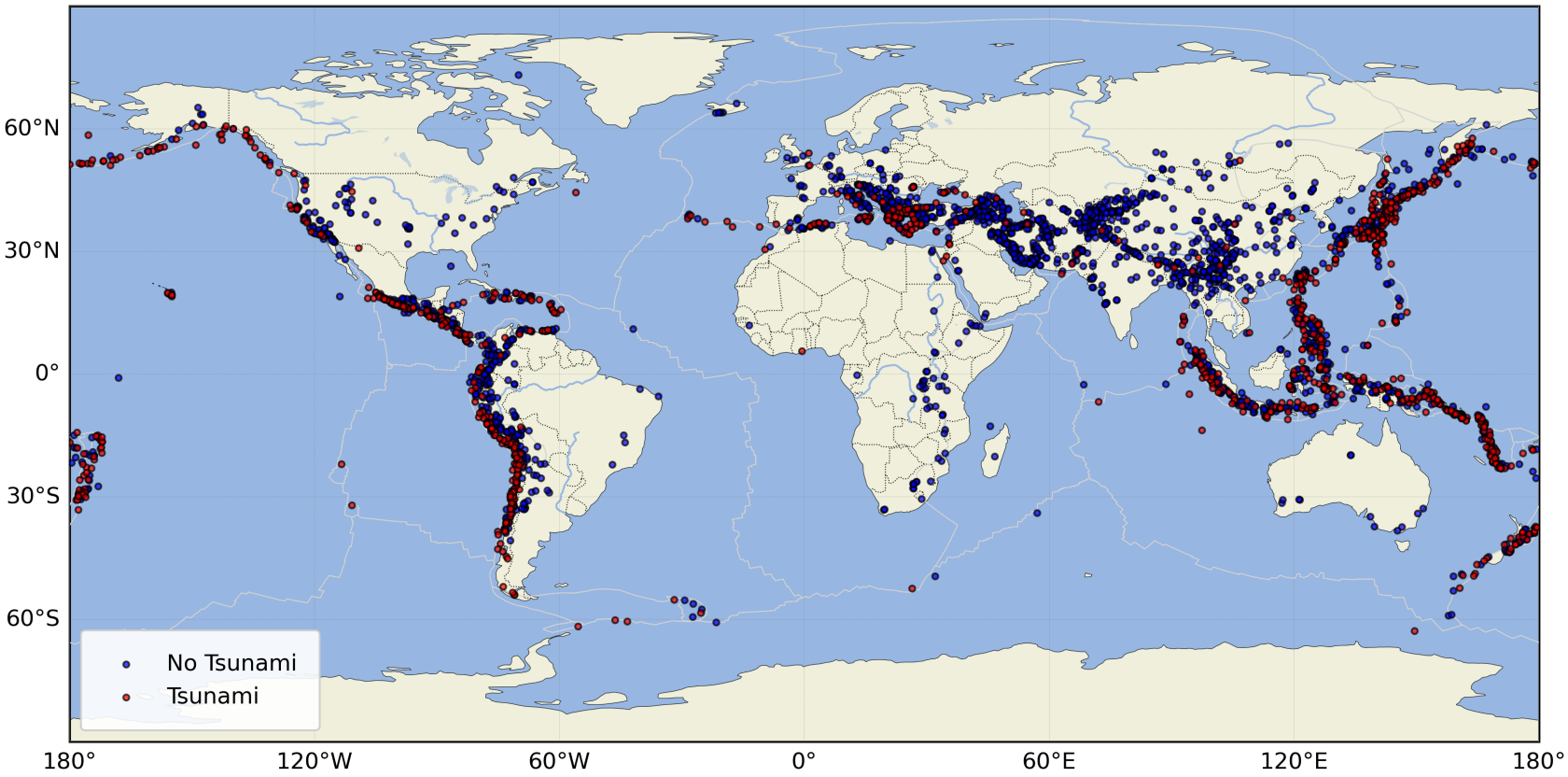
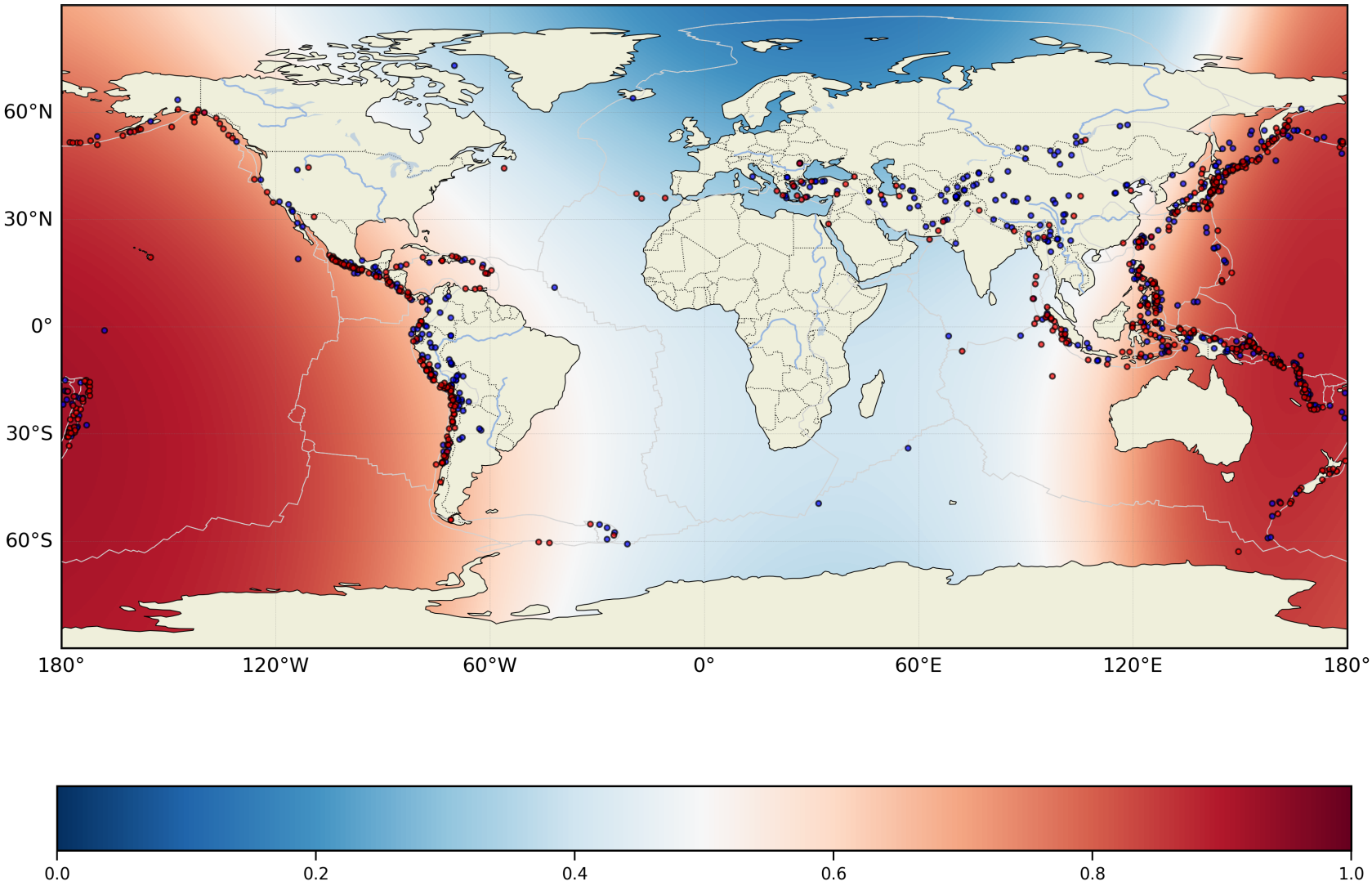


Figure: Point pattern of earthquakes (red) and associated tsunami occurrences (blue).

Real Data Illustration

POC Surface—Depth: Percentile 5



Part II

Generative Transformer-Based Approaches for Cascading Extremes

Rationale

Cascades are Natural Models for Generative Extremes

- Chains of extreme events and cascades of extreme events as proposed in de Carvalho et al.2026) offer a natural starting point for thinking about generative extremes.

Context

Attention is All you Need

- The starting point for the construction of our first generative AI approach for extreme events is based on notions of multi-headed attention as well as of transformers, as introduced in Vaswani et al (2017); for an introduction see Bishop & Bishop (2023; §12).

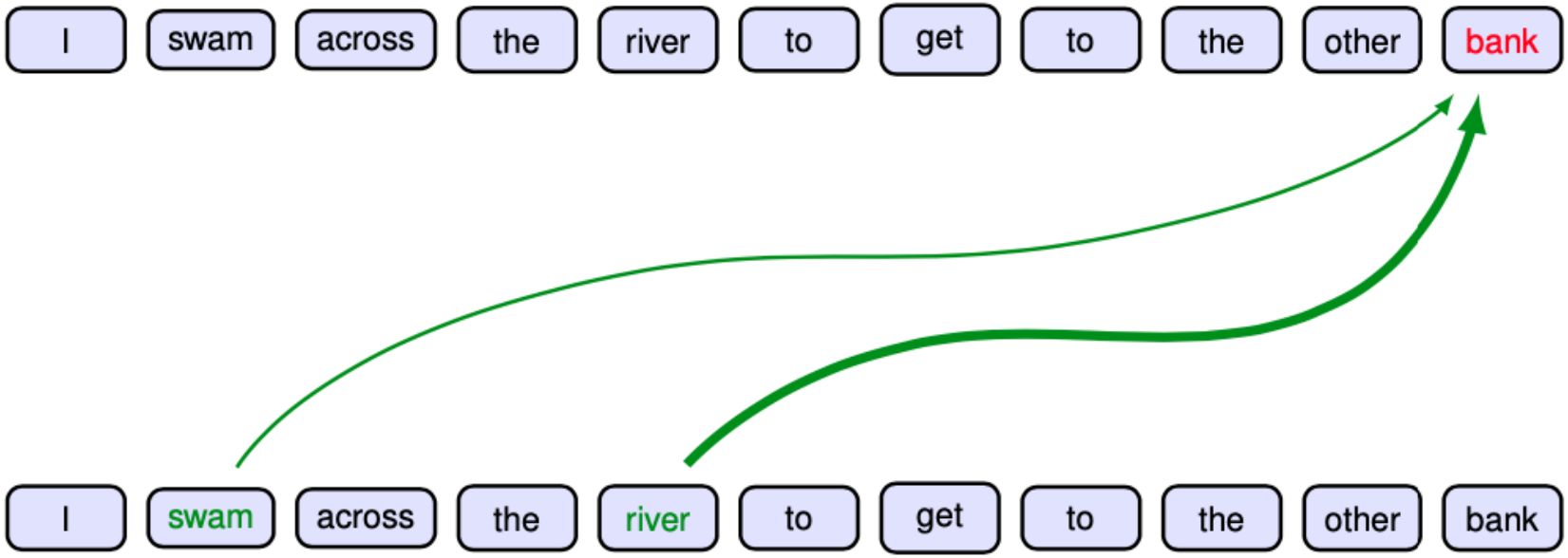
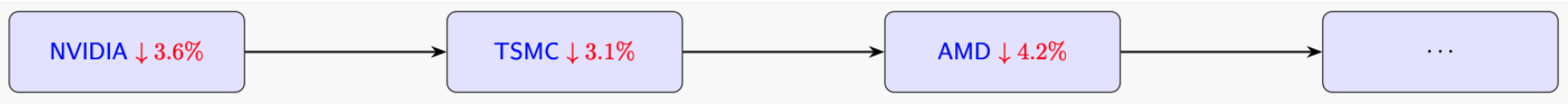


Figure 12.1 Schematic illustration of attention in which the interpretation of the word 'bank' is influenced by the words 'river' and 'swam', with the thickness of each line being indicative of the strength of its influence.

Context

- Whereas the unit of analysis in standard transformers is a sequence of tokens, in our EVT-based transformer is a sequence of extreme events.



- By analogy with language models, where the meaning of a token depends on its context, the interpretation of an extreme event similarly depends on the surrounding events.

Example

The sequence of extreme stock losses shown in the above figure is more indicative of a tech-specific cascade than of a broader macro-financial shock.

The proposed model is conceived to learn from data both:

- the most probable continuations of extreme event sequences;
- the probabilities of subsequent extremes.

Setup

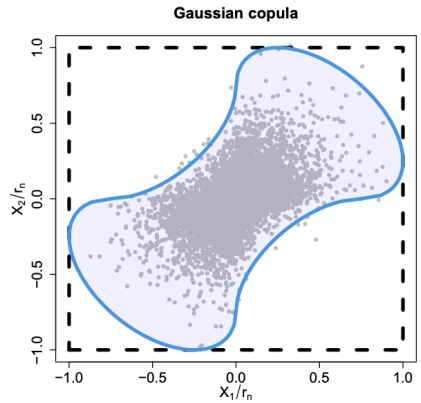
Introduction to Geometric Extremes

- Let $\{\mathbf{Y}_t\} = \{(Y_{1,t}, \ldots, Y_{d,t})^{\mathrm{T}}\}$, be a d-dimensional stochastic process with standard Laplace margins.
- Under regularity conditions (Nolde and Wadsworth2022; Wadsworth and Campbell2024; Murphy-Barltrop et al.2025) the joint density satisfies

$$-\log f_{\mathbf{Y}_t}(u\mathbf{x}) \sim u\, g_t(\mathbf{x}), \qquad \text{as } u \to \infty,$$

where the gauge function $g_t : \mathbb{R}^d \to \mathbb{R}_+$ is continuous and homogeneous, i.e., for any $c > 0$

$$g_t(c\mathbf{w}) = c g_t(\mathbf{w}).$$



- The sublevel set

$$G_t := \{\mathbf{x} \in \mathbb{R}^d : g_t(\mathbf{x}) \leq 1\}$$

Setup

Magnitude and Direction of Extremes

- Write $\mathbf{X}_t = R_t \mathbf{W}_t$ with

$$R_t = \|\mathbf{X}_t\|, \quad \mathbf{W}_t = \mathbf{X}_t / R_t \in \mathbb{S}^{d-1}.$$

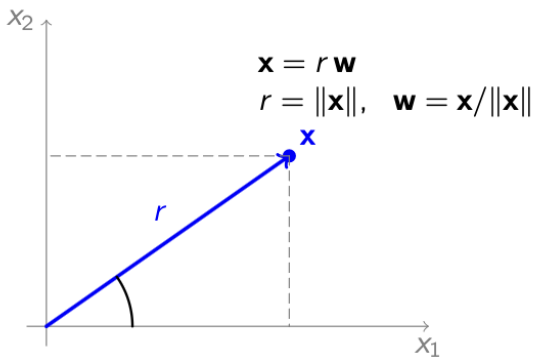
- On $\{R_t > u_t(\mathbf{w})\}$, the density of $\mathbf{X}_t$ factorizes

$$f_t(\mathbf{x}) = f_{R_t|\mathbf{W}_t}(r \mid \mathbf{w})\, h_t(\mathbf{w}), \qquad \mathbf{x} = r\mathbf{w}.$$

- Here, h_t is the angular surface on $\mathbb{S}^{d-1}$ (Castro et al.2018) and $f_{R_t|\mathbf{W}_t}$ is a truncated gamma density (Murphy-Barltrop et al.2025):

$$f_{R_t|\mathbf{W}_t}(r \mid \mathbf{w}) \propto \{g_t(\mathbf{w})\}^d\, r^{d-1} \exp\{-r g_t(\mathbf{w})\},$$

for $r > u_t(w)$.



Time Between Extremes

- Time between exceedances is modelled using a (marked) Hawkes process.
- Each exceedance temporarily raises the intensity (self-exciting).
- The marks take values in $\mathbb{S}^{d-1} \times \mathbb{R}_+$.
- The exceedance times $T_1 < T_2 < \cdots$ and marks $(\mathbf{W}_i, R_i)$ satisfying $R_i > u_i$ form a marked point process

$$N = \sum_{i \geq 1} \delta_{(T_i, W_i, R_i)},$$

with intensity

$$\lambda_t = \mu_t + \sum_{T_j < t} \psi_j(t).$$

- Here $\mu_t > 0$ is the baseline intensity and $\psi_j(t) \geq 0$ is the excitation kernel associated with event j, with $\psi_j(t) = 0$ for $t \leq T_j$ and $\int_{T_j}^{\infty} \psi_j(t)\, dt < \infty$.

Cascading Extremes

- For each event i, the parent variable $P_i \in \{0, 1, \ldots, i-1\}$ has distribution

$$\mathrm{P}(P_i = 0) = \frac{\mu_{T_i}}{\lambda_{T_i}}, \qquad \mathrm{P}(P_i = j) = \frac{\psi_j(T_i)}{\lambda_{T_i}}, \quad 1 \leq j < i.$$

- An event with $P_i = 0$ is an immigrant; an event with $P_i = j \geq 1$ is an offspring of event j.

Definition (Cascade)

A cascade $\mathcal{C}$ consists of an immigrant i_0 and all its descendants under the parent relation.

- Since $P_i < i$, each cascade is a rooted tree.
- The offspring set of j is

$$\nabla_j(\mathcal{C}) = \{i \in \mathcal{C} : P_i = j\}.$$

Cascading Extremes

- By standard theory of Hawkes processes, each event i generates offspring as a Poisson process on (T_i, ∞) with rate $\psi_i(t)$, independently across events.
- The mean number of offspring is the branching ratio

$$\nu_i = \int_{T_i}^{\infty} \psi_i(t)\, dt.$$

- Recall that the process is subcritical if $\nu := \sup_{i \geq 1} \nu_i < 1$, a.s.

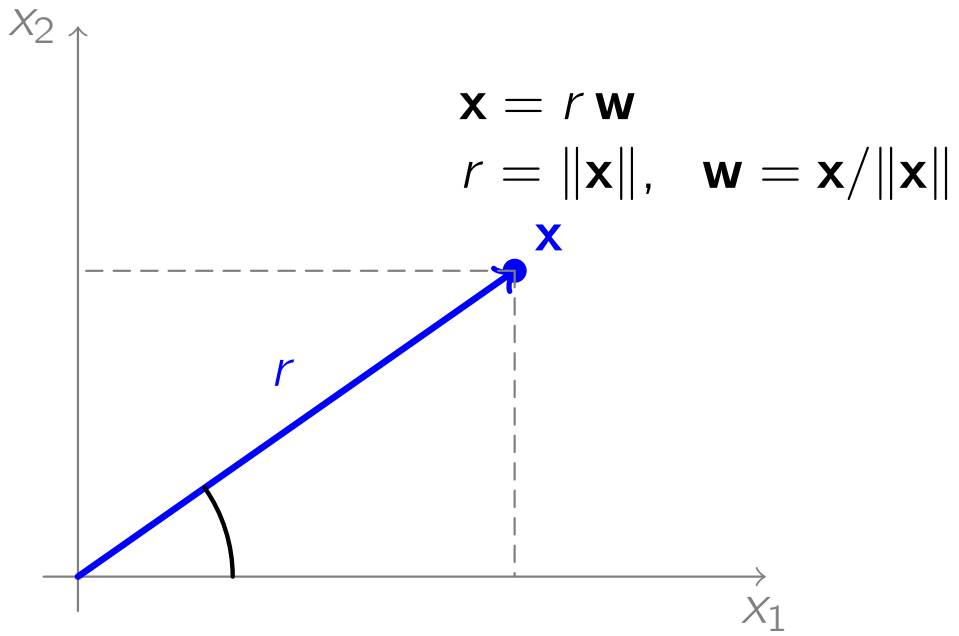
Theorem (Cascade termination)

Suppose $\nu < 1$. Then every cascade $\mathcal{C}$ satisfies:

(a) $|\mathcal{C}| < \infty$, *a.s.;*

(b) $\mathsf{E}(|\mathcal{C}|) \leq (1 - \nu)^{-1}$.

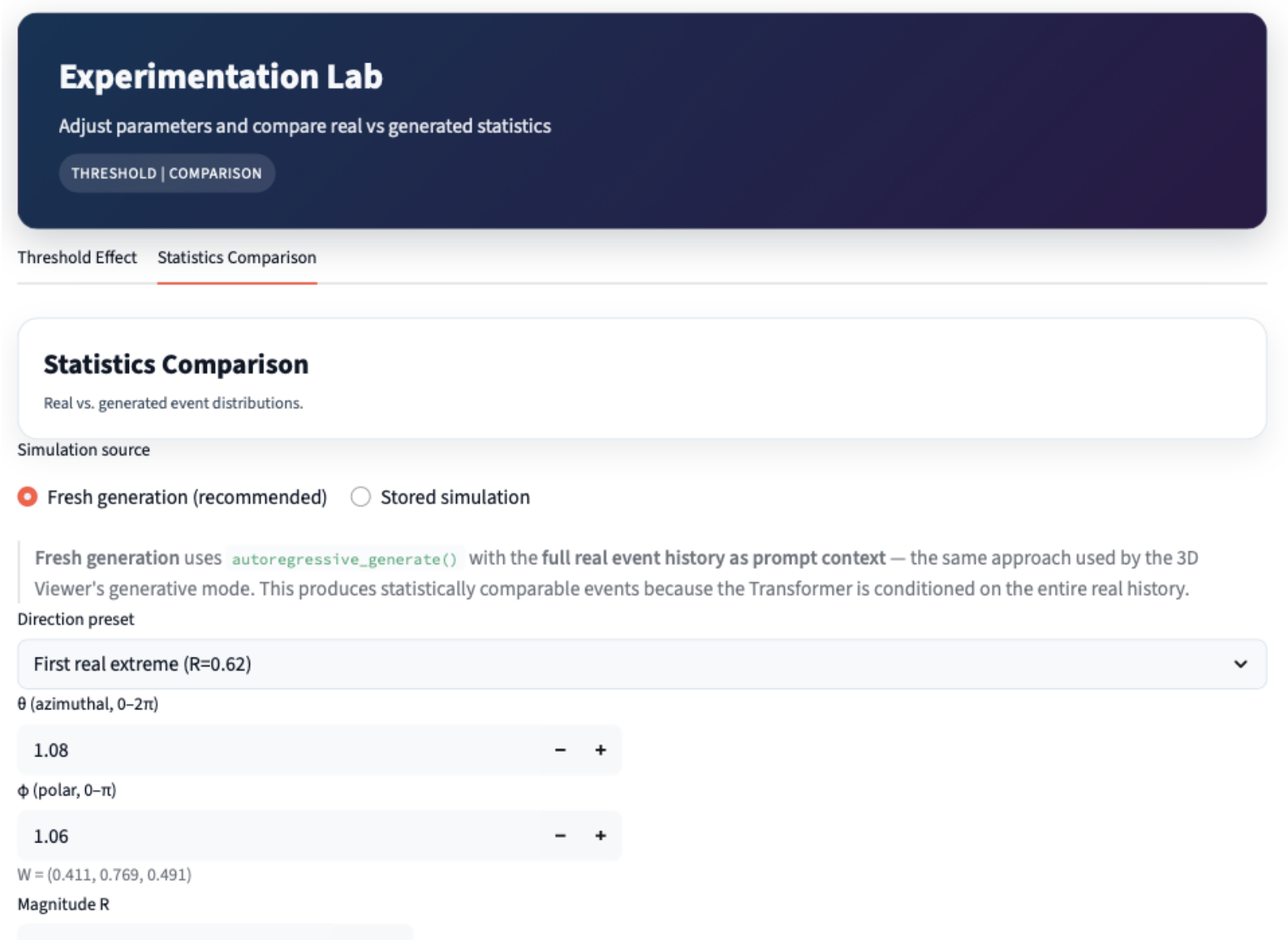
Summary of Model



- Angular component: Mixture of von Mises.
- Radial component: Truncated gamma (Wadsworth–Campbell).
- Time between exceedances: Hawkes process.

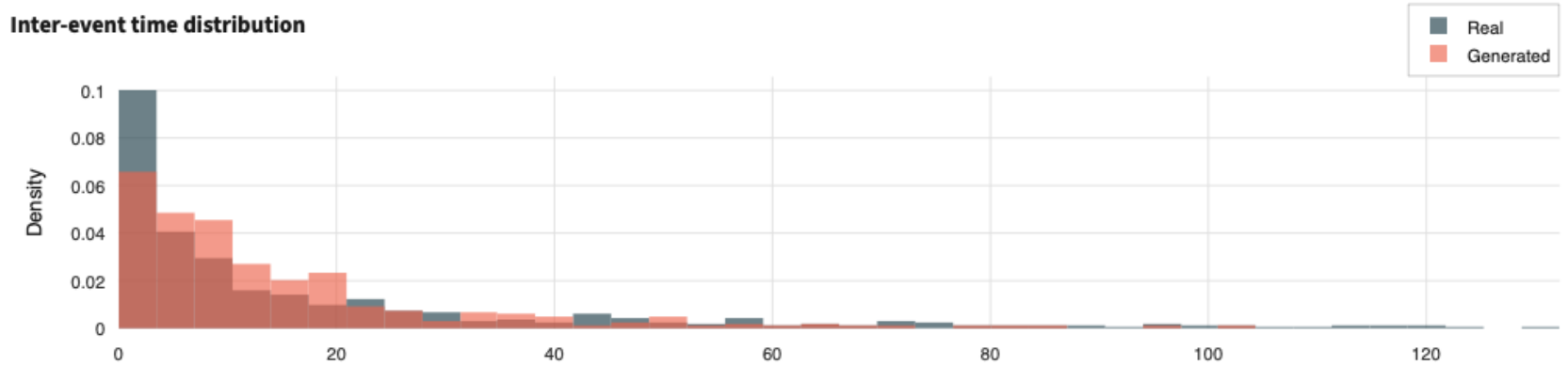
Backend of our Risk Enginge

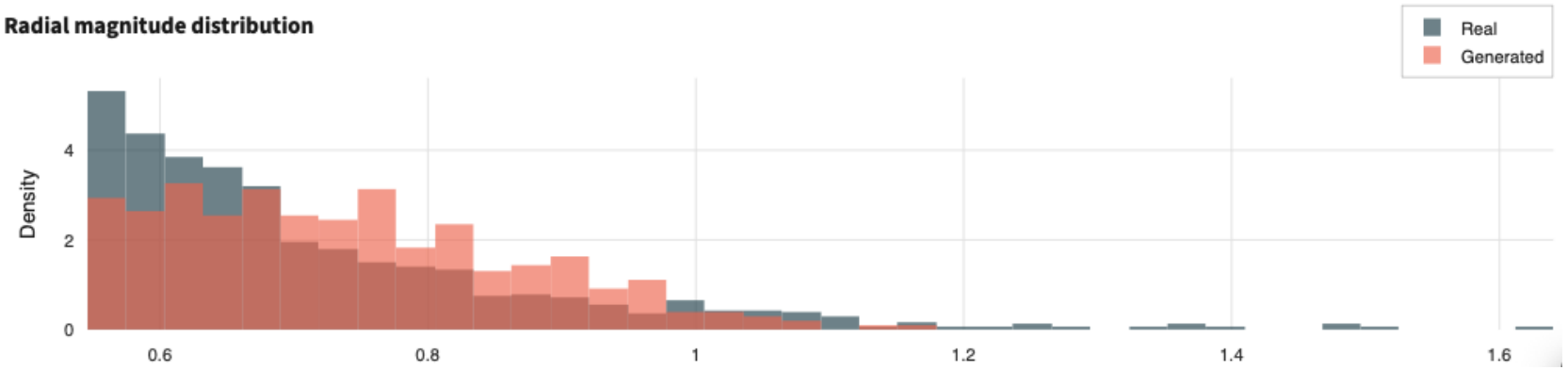
Experimentation Lab

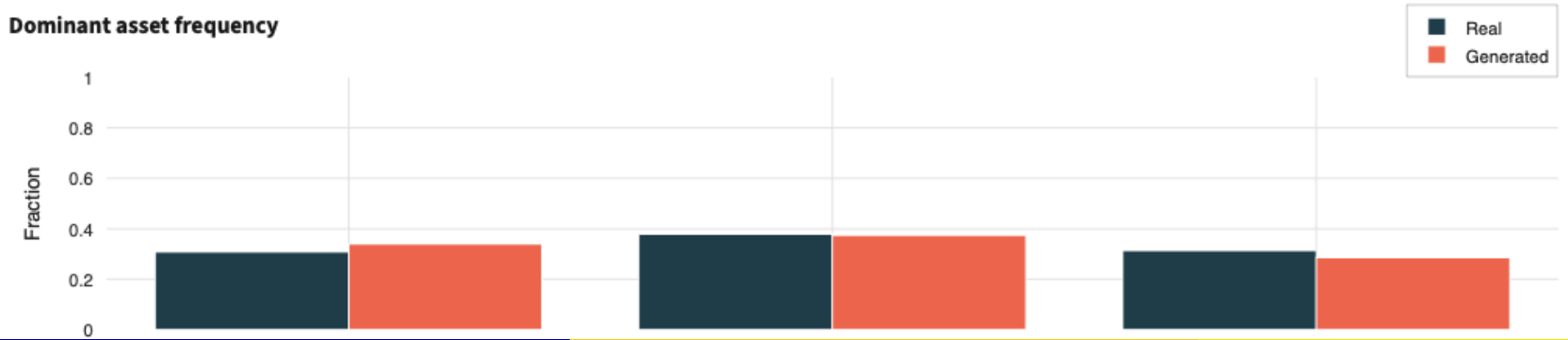


Backend of our Risk Enginge

Experimentation Lab







Closing Remarks

Summary

- This talk presented a novel statistical framework to tackle the rising concern of cascading extreme events—like tsunamis followed by earthquakes or heatwaves sparking wildfires, which in turn lead to further losses.
- The proposed approach aims to offer a novel outlook into triggering extremal events and their domino effects.
- KANE, a neural model based on Kolmogorov's superposition theorem, was developed to learn about the proposed POC surface.
- In addition, I offered some remarks on how cascades offer a natural starting point for thinking about generative extremes.
- Whereas the unit of analysis in standard transformers is a sequence of tokens, in our EVT-based transformer is a sequence of extreme events.

Basics on Transformers

Bishop and Bishop (2023, Ch. 12)

- Suppose that we are given a set of input tokens $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in an embedding space, and we wish to map them to a set of output tokens $\mathbf{y}_1, \ldots, \mathbf{y}_N$ that lies in a new embedding space that encodes a richer semantic structure.
- Basic idea:

$$\mathbf{y}_n = \sum_{m=1}^{N} a_{n,m} \mathbf{x}_m$$

where $a_{n,m} \geq 0$ and $\sum_{m=1}^{N} a_{n,m} = 1$ are are the so-called attention weights.

Further Details

Modus Operandi

Basic idea

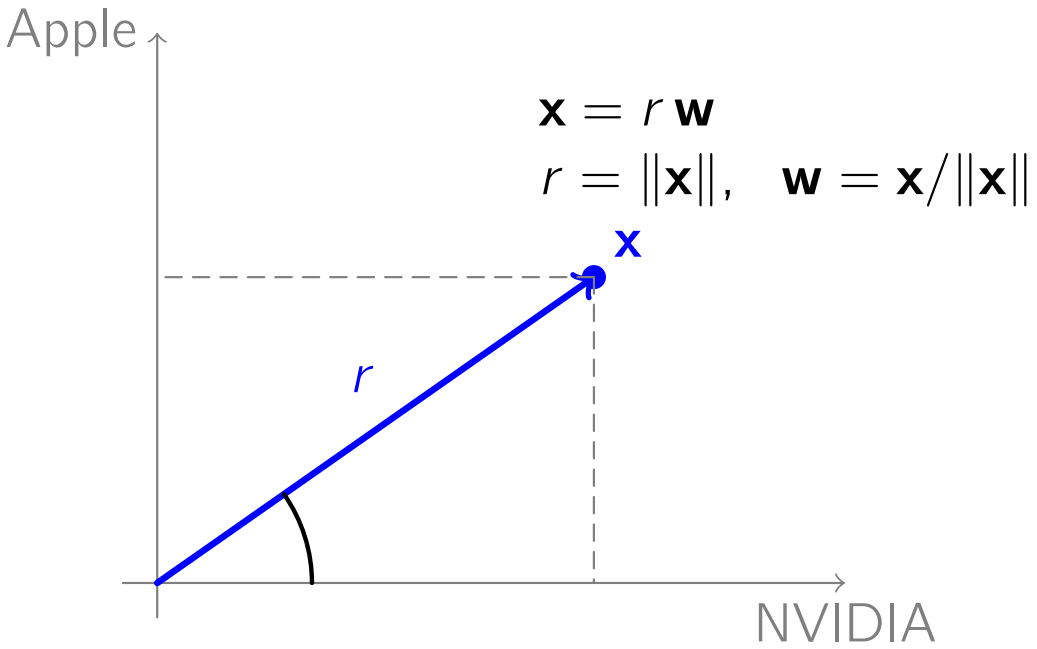
- Each event is $\mathbf{x}_i = (\mathbf{W}_i, R_i, \Delta T_i) \in \mathbb{R}^{d+2}$, where $\Delta T_i = T_i - T_{i-1}$.
- A transformer encoder maps the history until the ith period to a hidden state,

$$\mathbf{s}_i = f(\mathbf{x}_1, \ldots, \mathbf{x}_i) \in \mathbb{R}^D.$$

- Tree prediction heads (MLPs) then produce from $\mathbf{s}_i$ the parameters of the next-event densities. Loosely,

$$\mathbf{s}_i \mapsto \Theta_{i+1} = ((\pi, \boldsymbol{\mu}, \boldsymbol{\kappa}), \beta, \theta_{\text{Hawkes}}).$$

Further Details



- Angular component: $h(\mathbf{w}) = \sum_{k=1}^{K} f(\boldsymbol{\mu}_k, \boldsymbol{\kappa}_k)$.
- Radial component: $R \mid \mathbf{W} = \mathbf{w}, R > u_w \sim \text{TruncGamma}(d, g(\mathbf{w}))$.
- Time between exceedances: Hawkes process.

Modus Operandi

Basic idea

Transformer-Based Algorithm for Extremes

Input: trigger (W_0, R_0), horizon $t^* > 0$, trained encoder f_θ
$T_0 \leftarrow 0$, $\mathcal{H} \leftarrow \{(\mathbf{W}_0, R_0, T_0)\}$, $i \leftarrow 0$
while $T_i < t^*$ **do**
    $\mathbf{s}_i \leftarrow f(\mathbf{x}_0, \ldots, \mathbf{x}_i)$
    $\mathbf{W}_{i+1} \sim p(\mathbf{w} \mid \mathbf{s}_i)$
    $R_{i+1} \sim p(R \mid \mathbf{s}_i, \mathbf{W}_{i+1})$
    $\Delta T_{i+1} \sim p(\Delta T \mid \mathbf{s}_i)$
    $T_{i+1} \leftarrow T_i + \Delta T_{i+1}$
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\mathbf{W}_{i+1}, R_{i+1}, \Delta T_{i+1})\}$
    $i \leftarrow i + 1$
end while
return $\mathcal{H}$

Most Probable Continuation

- The most probable continuation can be defined as mode of the predictive distribution

$$\mathbf{x}^*_{n+1} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$$

  This is tantamout to greedy search (Bishop and Bishop2023).

- Evidently, this is not the same as the most probable sequence, which is given by maximizing the joint distribution

$$(\mathbf{x}^*_1, \ldots, \mathbf{x}^*_n) = \arg\max_{(\mathbf{x}_1, \ldots, \mathbf{x}_n)} p(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

Further Details

Implemented Model & Training

In practice

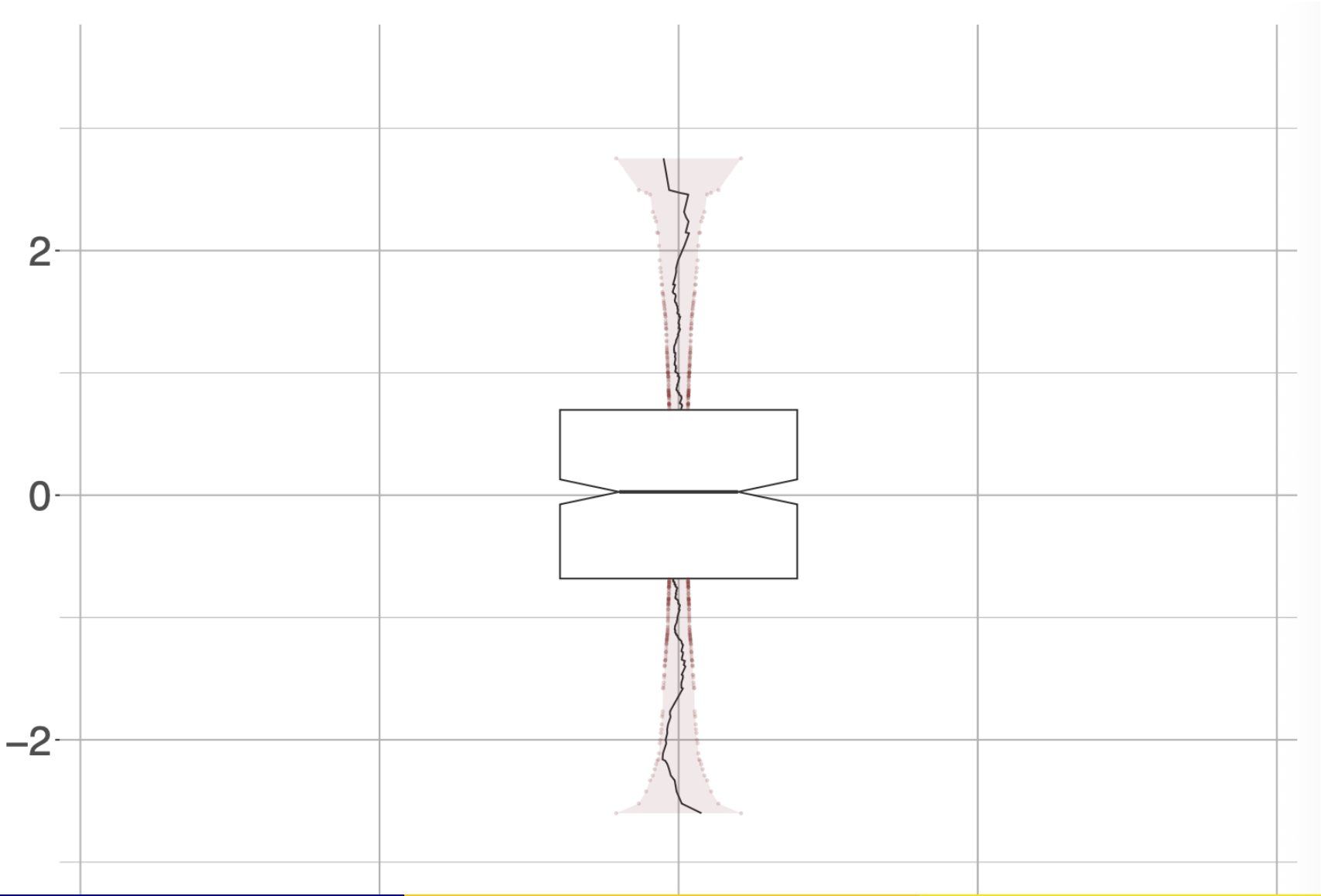
4 layers, $D = 128$, 4 attention heads per layer, sinusoidal positional enconding.

Training

- Adam optimizer, constant learning rate 10^4.
- Batch size 16, sequence length 128.
- 150 epochs, no early stopping, we save the best validation loss checkpoint.
- Trained on CPU.

Supporting Information

Model Checking



QQ Boxplot

Rodu and Kafadar (2022; *Journal of Computational and Graphical Statistics*)

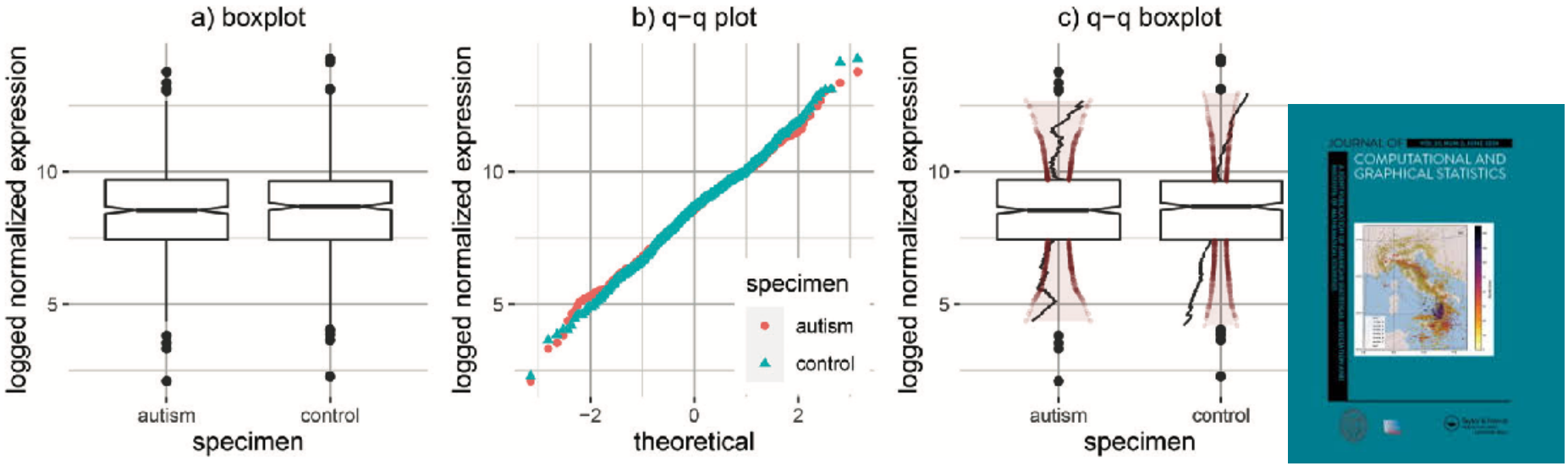


Figure 1. A comparison of the boxplot, q-q plot, and q-q boxplot highlights advantages of the q-q boxplot. Logged, normalized gene expression data for a patient with autism (left) and a "control" patient (right) as displayed by (a) boxplot; (b) q-q plot referenced to the normal distribution[3]; and (c) q-q boxplot referenced to the normal distribution. Data come from a random sample of the observations obtained from the Expression Atlas (Papatheodorou et al. 2019) (https://www.ebi.ac.uk/gxa/experiments/E-GEOD-30573/Results).

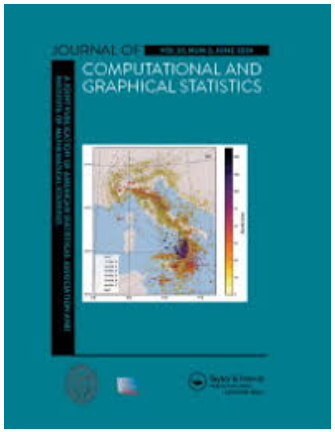
Randomized Quantile Residuals

Dunn and Smyth (1996; *Journal of Computational and Graphical Statistics*)

3. RANDOMIZED QUANTILE RESIDUALS

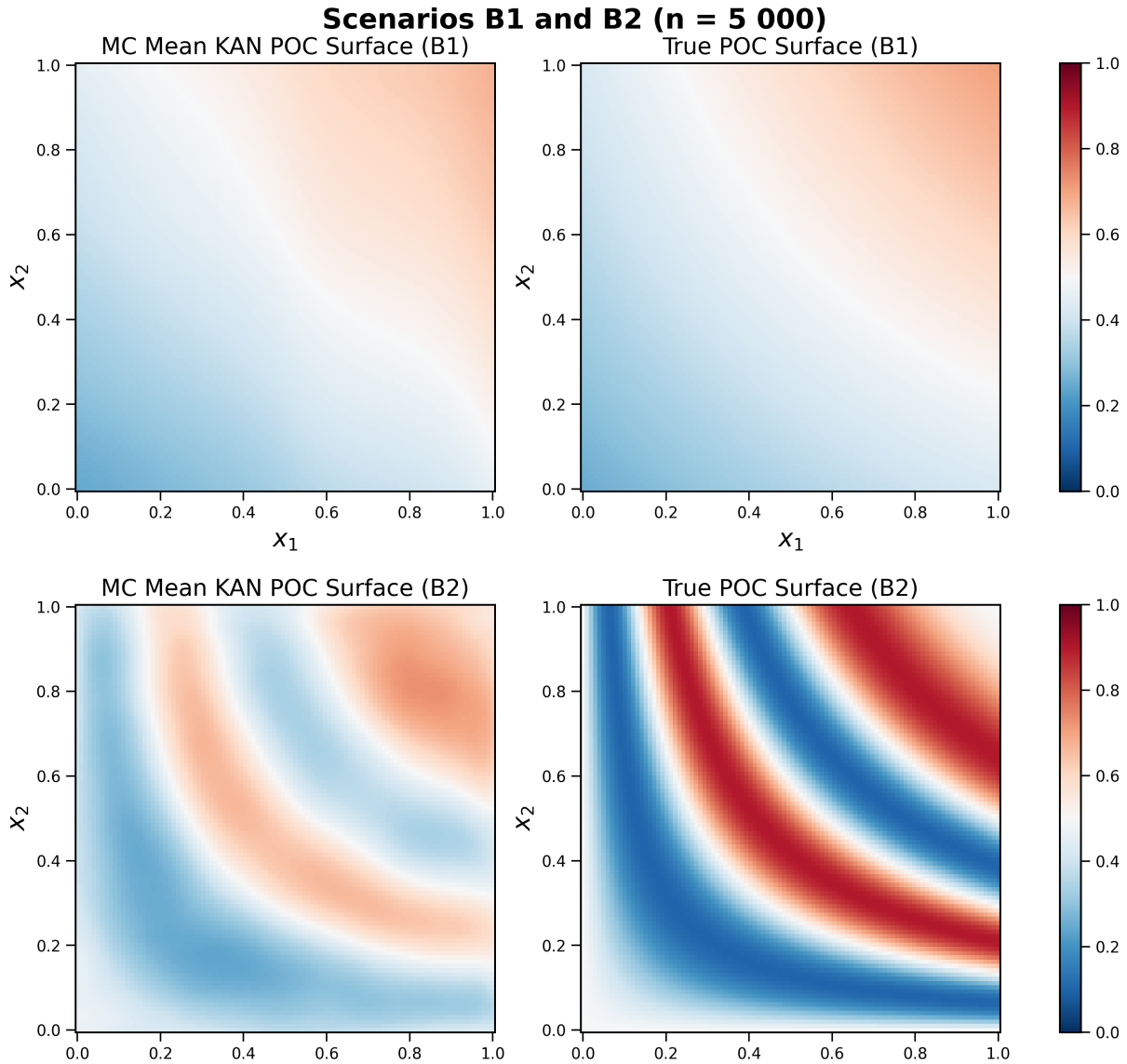
Let $F(y;\mu,\phi)$ be the cumulative distribution function of $\mathcal{P}(\mu,\phi)$. If F is continuous, then the $F(y_i;\mu_i,\phi)$ are uniformly distributed on the unit interval. In this case, the quantile residuals are defined by

$$r_{q,i} = \Phi^{-1}\{F(y_i;\hat{\mu}_i,\hat{\phi})\},$$



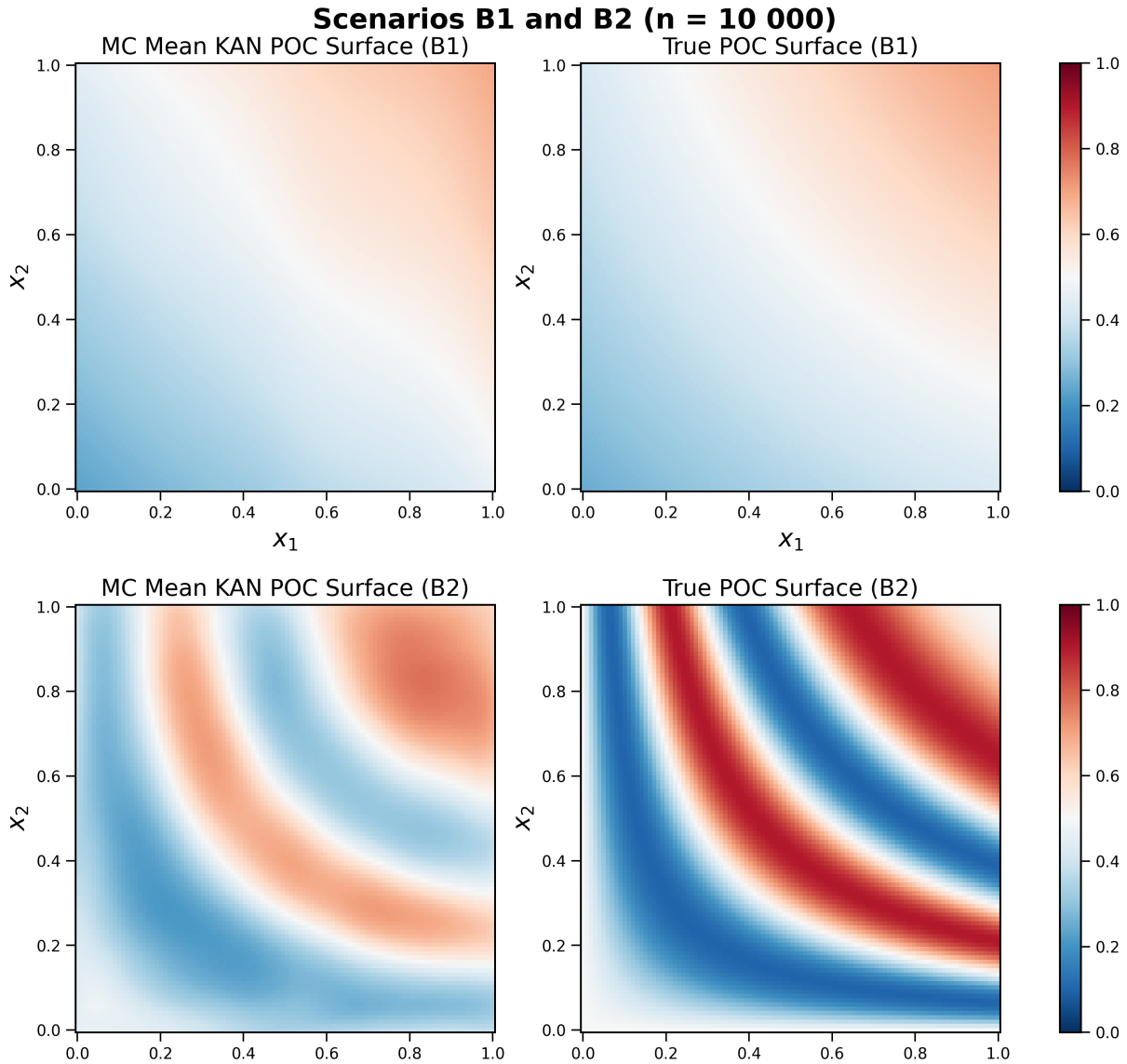
Randomized Quantile Residuals

Scenarios B



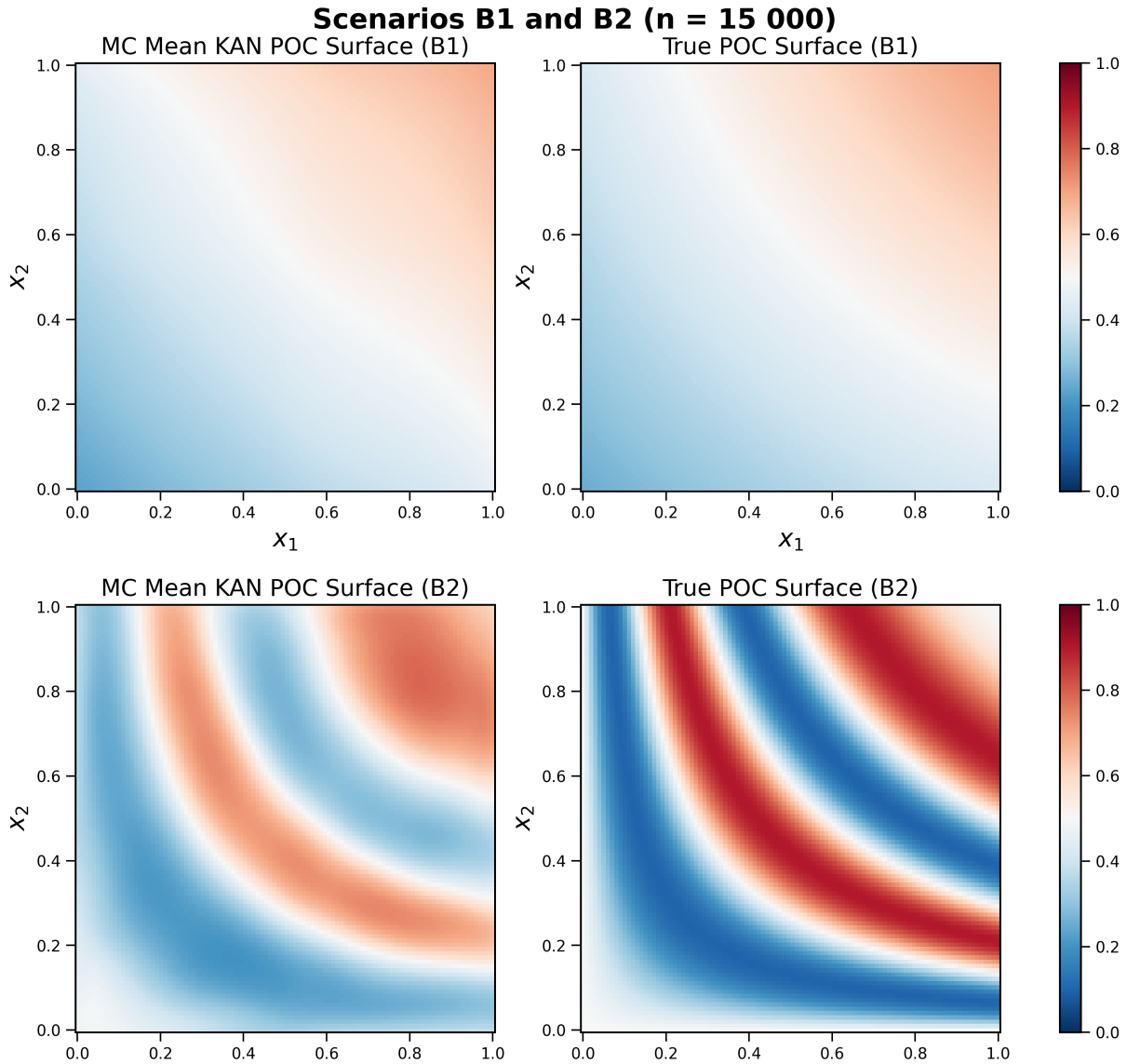
Randomized Quantile Residuals

Scenarios B



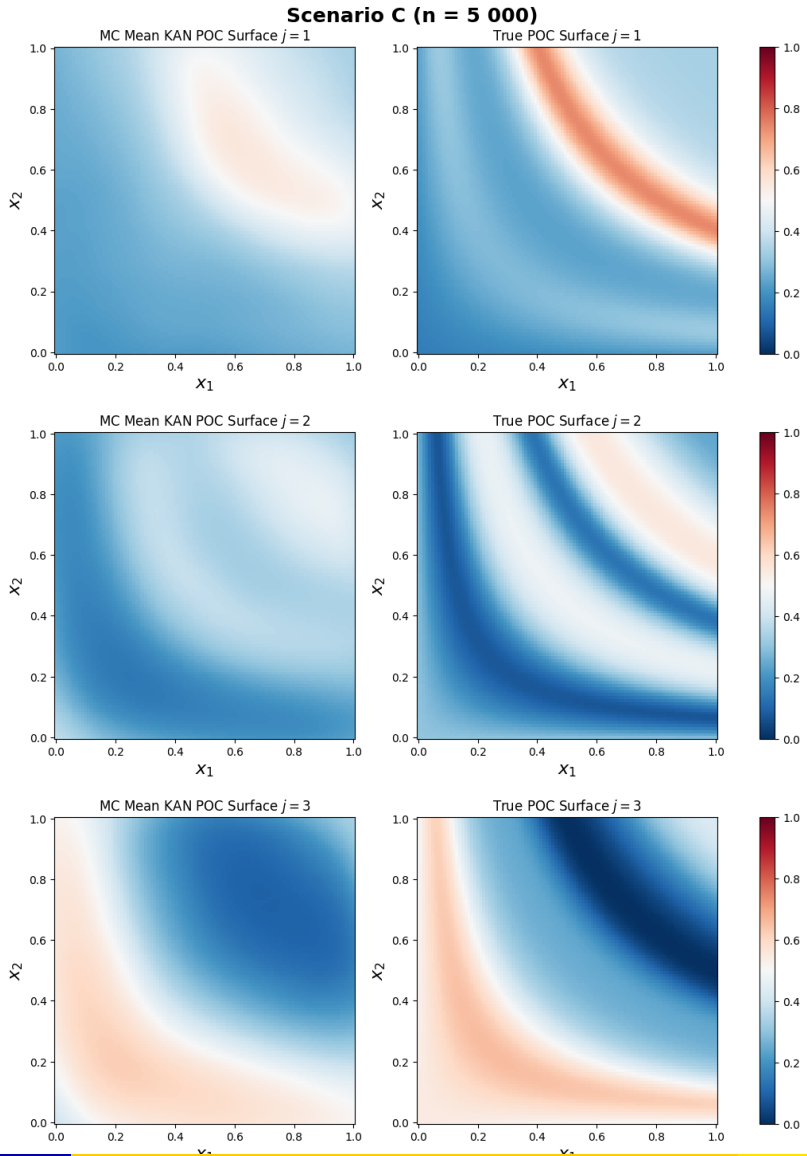
Randomized Quantile Residuals

Scenarios B



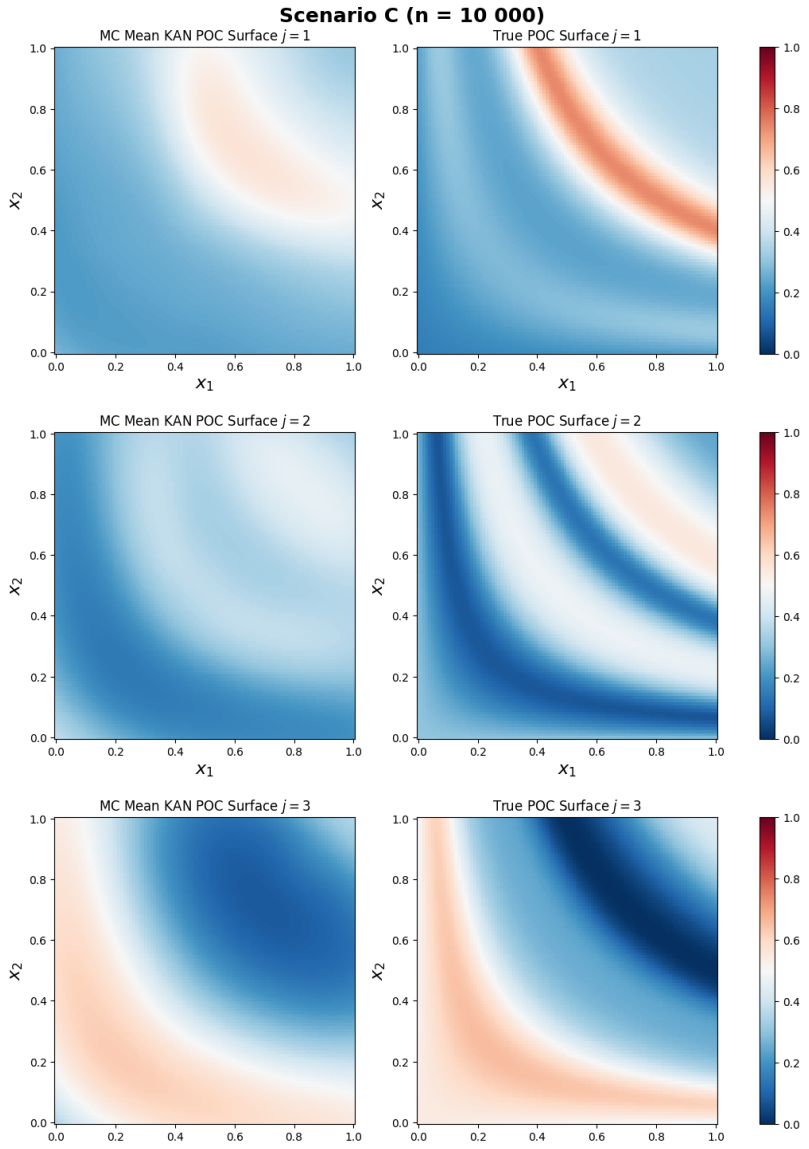
Randomized Quantile Residuals

Scenarios C



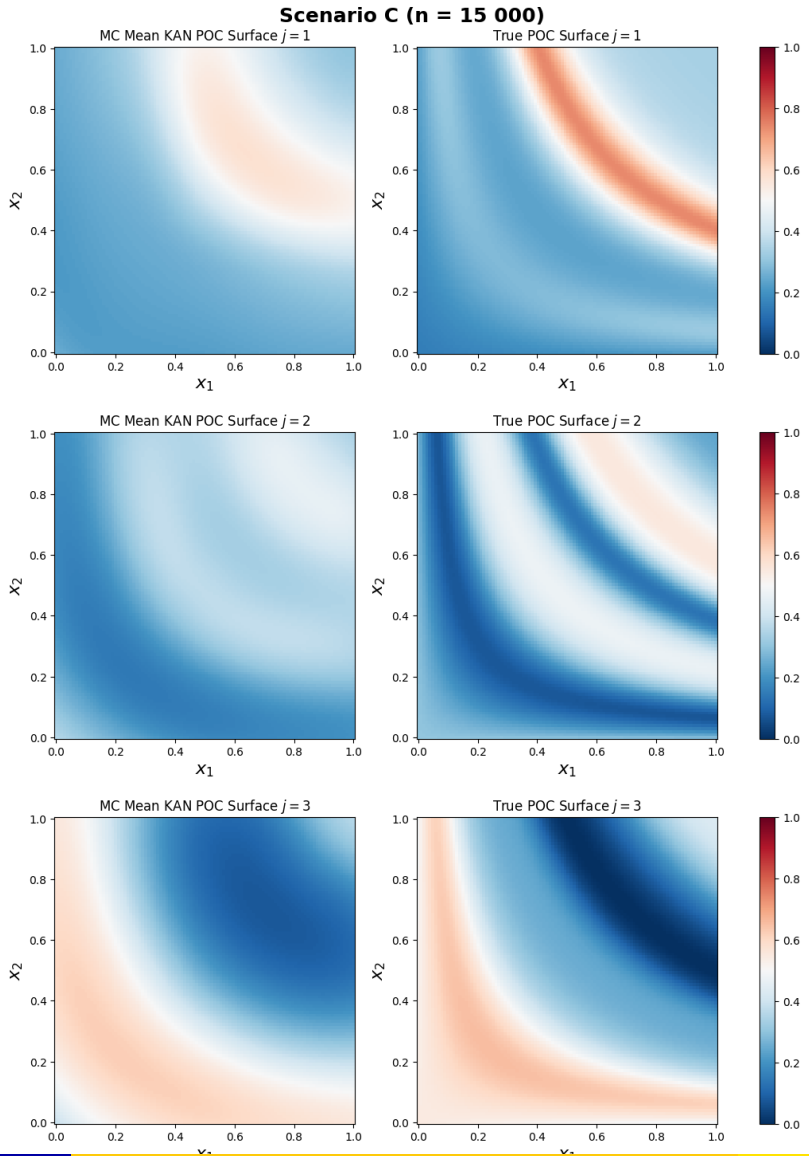
Randomized Quantile Residuals

Scenarios C



Randomized Quantile Residuals

Scenarios C



On the Continuity of POC

Theorem

Let $\{(I_{\mathbf{x},u}, Y_{\mathbf{x}}) : \mathbf{x} \in [0,1]^d,\ u > 0\}$. *Assume:*

(a) *The mapping* $\mathbf{x} \mapsto P(I_{\mathbf{x},u} = 1 \mid Y_{\mathbf{x}} > u)$ *is continuous on* $[0,1]^d$, *for all* $u \in \mathbb{R}$;

(b) *The limit* $\lim_{u \to y^*} P(I_{\mathbf{x},u} = 1 \mid Y_{\mathbf{x}} > u)$ *exists and convergence is uniform in* $\mathbf{x}$ *over* $[0,1]^d$.

Then, $\mathbf{x} \mapsto \alpha(\mathbf{x}) \equiv \lim_{u \to y^*} P(I_{\mathbf{x},u} = 1 \mid Y_{\mathbf{x}} > u)$ *is continuous on* $[0,1]^d$.

Universal Approximation Theorem

Theorem (Universal Approximation Theorem)

Suppose f is a continuous function on a compact space $\mathcal{X} \subset \mathbb{R}^d$ and σ is not a polynomial. Then, for any $\varepsilon > 0$, there exists a one-hidden layer neural network $f_{\boldsymbol{\theta}}$ such that

$$\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x})| < \varepsilon.$$

Notation: Here $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ is a one-hidden layer feedforward neural network composed of K neurons

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = b^{(2)} + \sum_{k=1}^{K} w_k^{(2)} \sigma(\langle \mathbf{w}_k^{(1)}, \mathbf{x} \rangle + b_k^{(1)}),$$

where

$$\boldsymbol{\theta} := \{b^{(2)}\} \cup \{\mathbf{w}_k^{(1)}, w_k^{(2)}, b_k^{(1)}\}_{k=1}^{K} \in \Theta := \mathbb{R} \times (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^K,$$

and where $\langle \cdot, \cdot \rangle$ is a scalar product on $\mathbb{R}^d$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function.

Berlyand, L. and Jabin, P.-E. (2023), *Mathematics of deep learning: An introduction*, Boston: de Gruyter.

Bishop, C. M. and Bishop, H. (2023), *Deep learning: Foundations and concepts*, New York: Springer.

Castro, D., de Carvalho, M., and Wadsworth, J. L. (2018), "Time-Varying Extreme Value Dependence with Application to Leading European Stock Markets," *Annals of Applied Statistics*, 12, 283–309.

Cutter, S. L. (2018), "Compound, cascading, or complex disasters: What's in a name?" *Environment: Science and Policy for Sustainable Development*, 60, 16–25.

de Carvalho, M., Ferrer, C., and Vallejos, R. (2026), "A Kolmogorov–Arnold Neural Model for Cascading Extremes," *Extremes*.

Fakhoury, D., Fakhoury, E., and Speleers, H. (2022), "ExSpliNet: An interpretable and expressive spline-based neural network," *Neural Networks*, 152, 332–346.

Lee, J., de Carvalho, M., Rua, A., and Avila, J. (2024), "Bayesian smoothing for time-varying extremal dependences," *Journal of the Royal Statistical Society, Ser. C*, 73, 581–597.

Lin, J.-N. and Unbehauen, R. (1993), "On the realization of a Kolmogorov network," *Neural Computation*, 5, 18–20.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. (2024), "KAN: Kolmogorov–Arnold networks," *arXiv:2404.19756*.

Montanelli, H. and Yang, H. (2020), "Error bounds for deep ReLU networks using the Kolmogorov–Arnold superposition theorem," *Neural Networks*, 129, 1–6.

Murphy-Barltrop, C., Wadsworth, J., de Carvalho, M., and Youngman, B. (2025), "Modelling non-stationary extremal dependence through a geometric approach," *arXiv:2509.22501*.

Nolde, N. and Wadsworth, J. L. (2022), "Linking representations for multivariate extremes via a limit set," *Advances in Applied Probability*, 54, 688–717.

Raymond, C., Horton, R. M., Zscheischler, J., Martius, O., AghaKouchak, A., Balch, J., Bowen, S. G., Camargo, S. J., Hess, J., Kornhuber, K., et al. (2020), "Understanding and managing connected extreme events," *Nature Climate Change*, 10, 611–621.

Sprecher, D. A. and Draghici, S. (2002), "Space-filling curves and Kolmogorov superposition-based neural networks," *Neural Networks*, 15, 57–67.

Thas, O., Neve, J. D., Clement, L., and Ottoy, J.-P. (2012), "Probabilistic index models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74, 623–671.

Wadsworth, J. L. and Campbell, R. (2024), "Statistical inference for multivariate extremes via a geometric approach," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86, 1243–1265.