Deep extremal regression: Grey box models for univariate and multivariate extremes

Jordan Richards

School of Mathematics and Maxwell Institute for Mathematical Sciences,
University of Edinburgh, UK

STOR-i 2026

Extremes

Any event, environmental or otherwise, that occurs with **very low** probability.

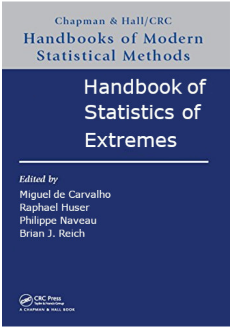
   

Outline

1. Give a brief introduction to extremal regression
2. Show how deep-learning can enhance modelling, in two settings:
    - **Marginal extremes:** Deep high-dimensional extremal quantile regression
      Joint with Raphaël Huser
    - **Multivariate extremal dependence:** Radial-angular modelling of multivariate extremes
      Joint with Ed Mackay, Callum Murphy-Barltrop, Phil Jonathan

These models are grey box – **Principled statistical models** combined with **data-driven deep learning methods**.

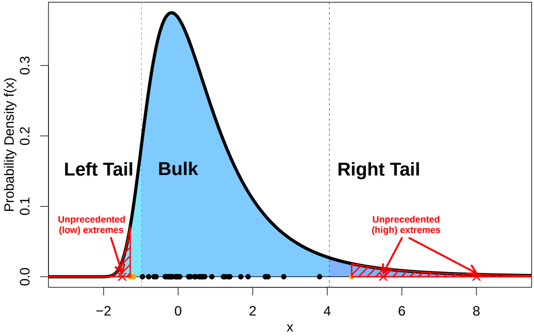
Outline



Chapter 21. Richards and Huser (2026a).
Extreme Quantile Regression with Deep Learning.
Code and short courses on GitHub - `https://github.com/Jbrich95/`

Background on extreme events

- Univariate context: Observations $Y_1, \ldots, Y_n$.
  Can we estimate the probability of unprecedented extreme events of a given size (typically larger than $M_n = \max(Y_1, \ldots, Y_n)$)?

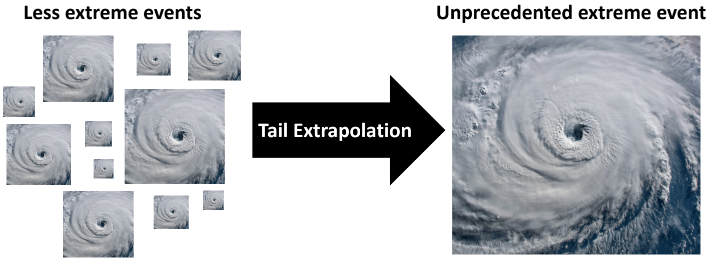


- Need to extrapolate into the tail, based on resilient, reliable, theoretically-justified methods

Main idea of Extreme-Value Theory (EVT)

Learn about unprecedented extreme events from less extreme events.

Requires assumptions about the regularity of the tail, or the tail decay rate



Marginal modeling of extremes — theory for peaks-over-threshold

- **Pickands–de Haan–Balkema Theorem:** high threshold excesses $Y - u \mid Y > u$ may be approximated by the generalized Pareto (GP) distribution, in the sense that there exists a scaling function $a(u) > 0$ such that as $u \to y_F$ (upper endpoint)

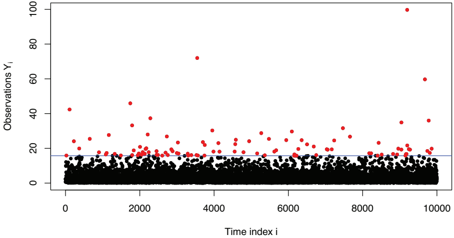
$$\Pr\left(\frac{Y-u}{a(u)} > y \mid Y > u\right) \to 1 - H(y) := \begin{cases} (1 + \xi y/\tau)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-y/\tau), & \xi = 0, \end{cases},$$

where $\tau > 0$ and $\xi \in \mathbb{R}$ are GP scale and shape parameters, respectively.

- In practice, we can model excesses directly as $Y - u \mid Y > u \sim \mathrm{GP}(\tau, \xi)$.

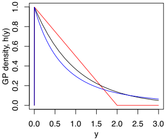
Threshold exceedance approach

$Y_1, Y_2, \ldots, \overset{\text{iid}}{\sim} F$



Pickands–de Haan–Balkema Theorem: For a broad range of distributions F, we have the following large-u approximation

$$Y_i - u \mid Y_i > u \sim \mathrm{GP}(\tau, \xi).$$



Extremal regression

Now, introduce data pairs $(Y_i, \mathbf{X}_i)_{i=1}^n$.

Stats 101

Given Y and covariates $\mathbf{X} \in \mathbb{R}^p$, how do we model the impacts of $\mathbf{X}$ on the extremes of Y?

Answer: **Regression**

In the GP case, we model

$$Y_i - u(\mathbf{x}_i) | (Y_i > u(\mathbf{x}_i), \mathbf{X} = \mathbf{x}_i) \sim \mathrm{GP}(\tau(\mathbf{x}_\mathrm{i}), \xi(\mathbf{x}_\mathrm{i})).$$

Extremal regression

To model extremes of $Y \mid \mathbf{X} = \mathbf{x}$, assume $Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{F}(\boldsymbol{\theta}(\mathbf{x}))$ with parameters $\boldsymbol{\theta} : \mathbb{R}^p \mapsto \mathbb{R}^q$.

Standard choices for $\mathcal{F}$ (in extreme analyses):

- $\mathcal{F} = \text{GP}$ with $\boldsymbol{\theta}(\mathbf{x}) = (\tau(\mathbf{x}), \xi(\mathbf{x}))$ (for threshold excesses);
- $\mathcal{F} = \text{GEV}$ with $\boldsymbol{\theta}(\mathbf{x}) = (\mu(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x}))$ (for maxima).
- *Newer* alternatives:
  - eGPD (Papastathopoulos and Tawn, 2013; Cisneros et al., 2024),
  - point processes (Richards and Huser, 2026b),
  - bGEV (Castro-Camilo et al., 2022; Richards and Huser, 2026a),
  - bGP (Majumder and Richards, 2025).

Majumder, Richards (2025). Semi-parametric bulk and tail regression using spline-based neural networks. arXiv.

Regression

To model extremes of $Y \mid \mathbf{X} = \mathbf{x}$, assume $Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{F}(\boldsymbol{\theta}(\mathbf{x}))$ with parameters $\boldsymbol{\theta} : \mathbb{R}^p \mapsto \mathbb{R}^q$.

Standard choices for $\boldsymbol{\theta}(\mathbf{x})$ (in extreme analyses):

- Linear: i.e., $\theta_j(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$.
- Additive (typically utilising smoothing splines): i.e., $\theta_j(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \sum_{k=1}^{K_i} \beta_{ik}\psi_{ik}(x_i)$.

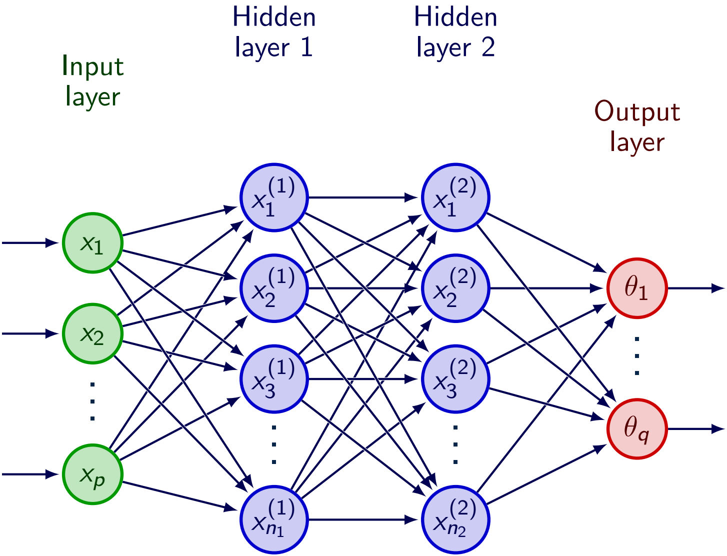
Recent uptick in replacing $\boldsymbol{\theta}(\mathbf{x})$ with a non-parametric alternative from supervised ML, e.g., **Neural Networks**:

$$\boldsymbol{\theta}(\mathbf{x}) = \mathbf{m}_1 \circ \mathbf{m}_2 \circ \cdots \circ \mathbf{m}_L(\mathbf{x}),$$

but this idea goes back to Carreau and Bengio (2007)!

Carreau, Julie and Bengio, Yoshua. (2007) A Hybrid Pareto Model for Conditional Density Estimation of Asymmetric Fat-Tail Data. AIaS.

Deep regression for extremes



Extremal regression – Inference

Deep extremal regression:

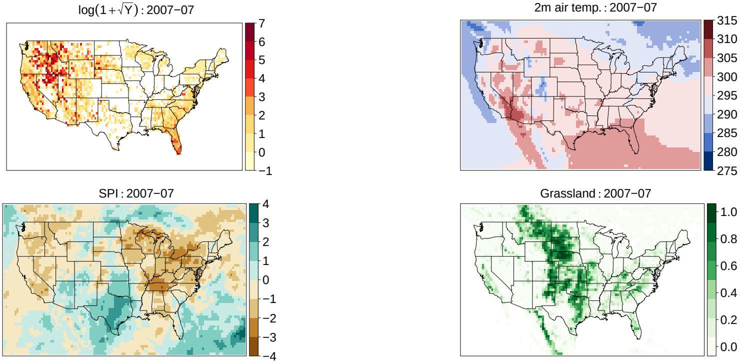
- Models fitted using variants of **stochastic gradient descent** with negative log-likelihood (associated with $\mathcal{F}$) as loss;
- Standard tools for regularisation, architecture optimisation;
- Different choices for the *type* of neural network - MLP, CNNs (Richards et al., 2023), LSTMs (Pasche and Engelke, 2024), GNNs (Cisneros et al., 2024);
- Can exploit off-the-shelf interpretability metrics - e.g., Shapley values, accumulated local effects;
- Can mix-and-match neural networks and interpretable functions - Partially-interpretable NNs (PINNs; Richards and Huser, 2026b).

Richards, Huser, Bevacqua, Zscheischler (2023). Insights into the drivers and spatiotemporal trends of extreme mediterranean wildfires with statistical deep learning. AIES.

Cisneros, Richards, Dahal, Lombardo, Huser (2024). Deep graphical regression for jointly moderate and extreme Australian wildfire. Spatial Statistics.

US Wildfire Extremes

- EVA 2021 Data Challenge - US Wildfire extremes in-fill
- Response Y burnt area from the Fire Program Analysis (Short, 2017)
- 1993–2015, Mar.–Sep., leaving 161 fields; $\approx$ 3500 locations; $p = 42$ predictors; $n = 216713$ (non-zero).



US Wildfire Extremes

PINN Model (Richards and Huser, 2026b):

- $\mathcal{F}$ taken as an extremal point process model with $q = 3$ parameters: focus on the "median" $q_\alpha(\mathbf{x})$ and shape $\xi(\mathbf{x})$.
- We let

$$\log q_\alpha(\mathbf{x}) = \mathbf{m}(\mathbf{x}) + \beta_1(\mathrm{State}) \cdot \mathrm{temperature} + \beta_2(\mathrm{State}) \cdot \mathrm{SPI},$$

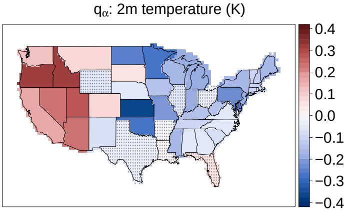
  where $\mathbf{m}(\cdot)$ is a CNN.
- $\xi(\mathbf{x}) := \mathrm{sigmoid}\{\xi((\mathrm{Lat}, \mathrm{Lon}))\}$ is an MLP with only spatial information.

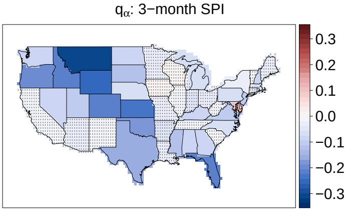
Model selection and training done via a leave-space-time-clusters-out validation (Appendix) and UQ via bootstrap.

Richards, Huser (2026). Regression modelling of spatiotemporal extreme US wildfires via partially-interpretable neural networks. JCGS.

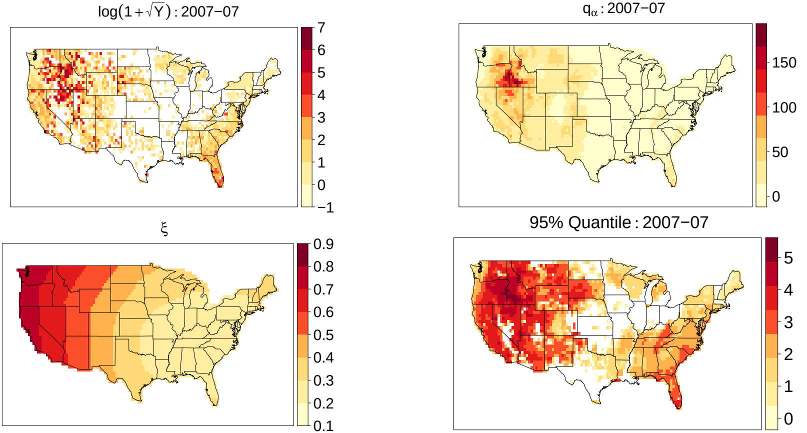
US Wildfire Extremes

State-varying coefficient model estimates:





US Wildfire Extremes



Metocean multivariate extremes

Now interested in joint extremes of a multivariate random vector $\mathbf{X} \in \mathbb{R}^d$.

- 31-year hindcast of hourly values for a site in Celtic sea ($n \approx 270000$).
- $d = 5$, x- and y-components of wind speed and wave height, and (log)-wave period

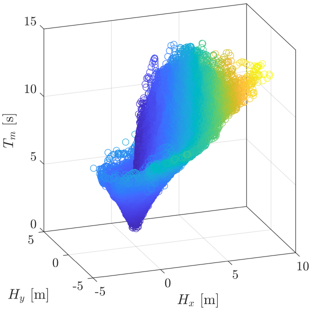
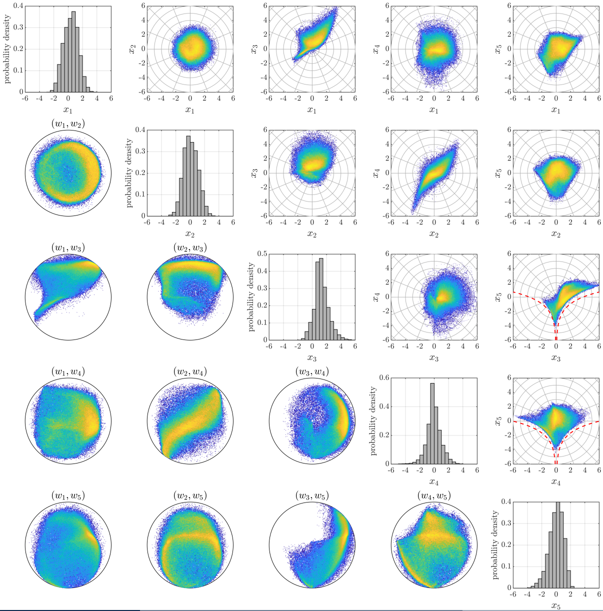


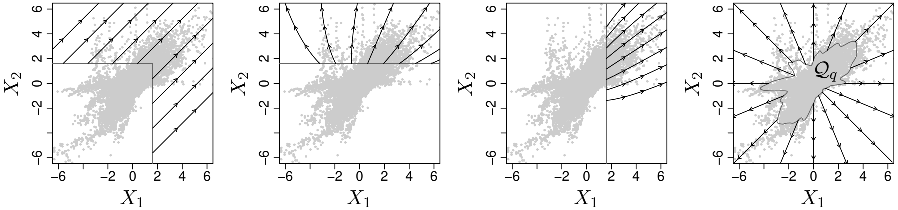
Figure: x-y components of windspeed H_s and T_m. Colours indicate the value of windspeed $H_s = \sqrt{H_x^2 + H_y^2}$ (low, high).

Metocean multivariate extremes



Multivariate extremes

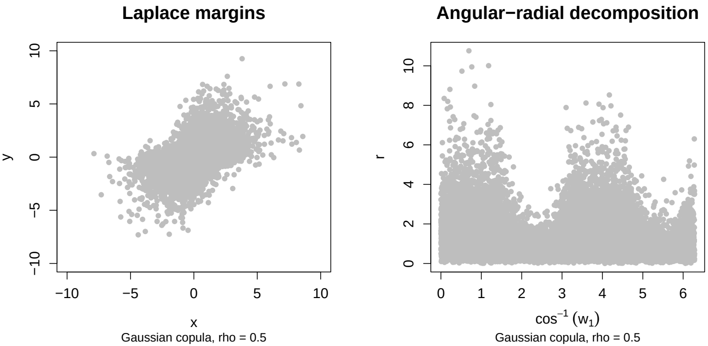
Ocean storms driven by non-convex combinations of "extremes" $\Rightarrow$ we model all joint extremes, rather than one variable being large.



Angular-radial representation

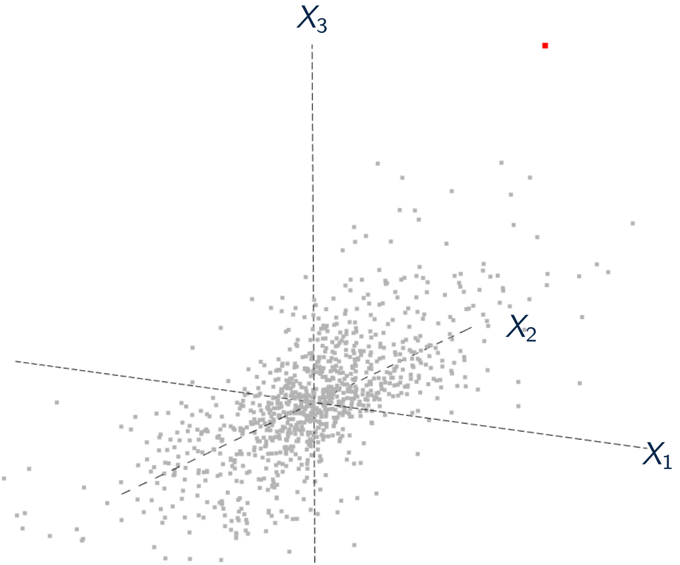
Let $\mathbf{X} \in \mathbb{R}^d$ and

$$R := \|\mathbf{X}\| > 0, \qquad \mathbf{W} := \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \mathcal{S}^{d-1}.$$

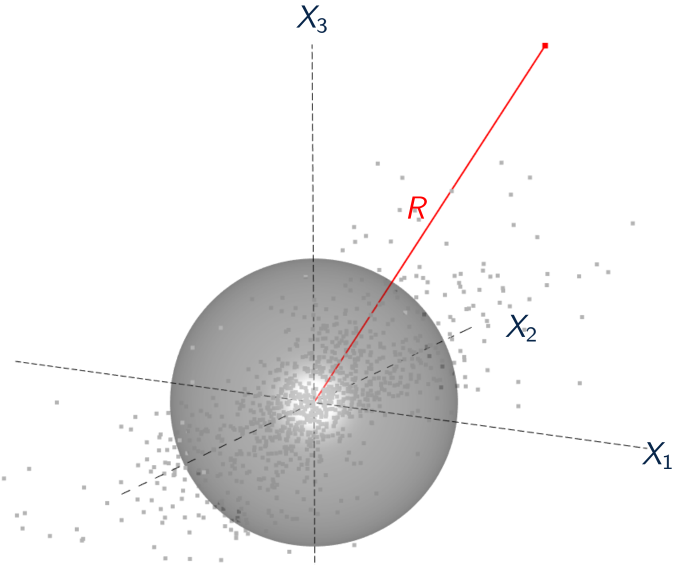


Murphy-Barltrop, C. J., Majumder, R., & Richards, J. (2024). Deep learning of multivariate extremes via a geometric representation. arXiv:2406.19936.

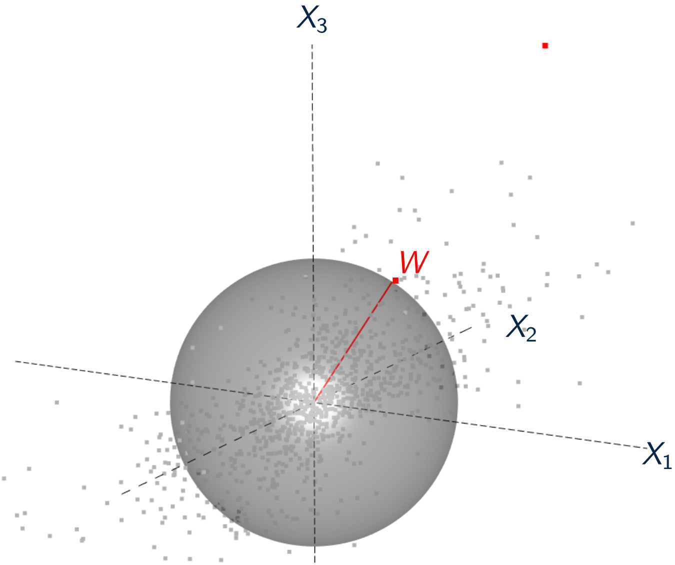
Angular-radial representations



Angular-radial representations



Angular-radial representations



Angular-radial representations

The joint density of $(R, \mathbf{W})$ can be written as

$$f_{\mathbf{R},\mathbf{W}}(r, w) = f_{\mathbf{W}}(\mathbf{w}) f_{R|\mathbf{W}}(r|\mathbf{w}). \tag{1}$$

The Semi-Parametric Angular-Radial model (SPAR; Mackay and Jonathan, 2023) replaces $f_{R|\mathbf{W}}(r|\mathbf{w})$ with a conditional GP model **when r is big**, i.e.,

$$f_{R|\mathbf{W}}(r|\mathbf{w}) \approx \zeta f_{\mathrm{GP}}(r - u(\mathbf{w}); \tau(\mathbf{w}), \xi(\mathbf{w})), \tag{2}$$

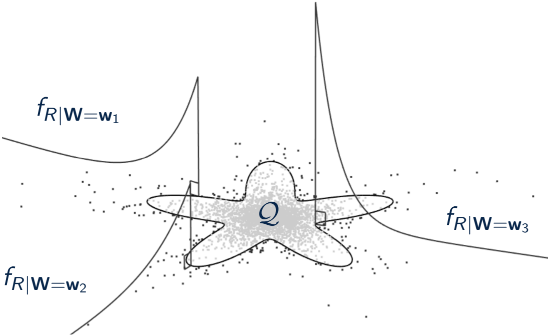
for large $r > u(\mathbf{w})$ and where $\zeta := \Pr(R > u(\mathbf{w})|\mathbf{W} = \mathbf{w})$ is fixed.

- Mackay et al. (2025) models $(u(\mathbf{w}), \tau(\mathbf{w}), \xi(\mathbf{w}))$ via **deep extremal regression**.

Mackay, Jonathan (2023). Modelling multivariate extremes through angular-radial decomposition of the density function. arXiv.

Mackay, Murphy-Barltrop, Richards, Jonathan (2025). Deep Learning Joint Extremes of Metocean Variables Using the SPAR Model. JOMAE.

Angular-radial representations



De Monte, L., Huser, R., Papastathopoulos, I., & Richards, J. (2025). Generative modelling of multivariate geometric extremes using normalising flows. arXiv:2505.02957.

Angular-radial representations

Recall

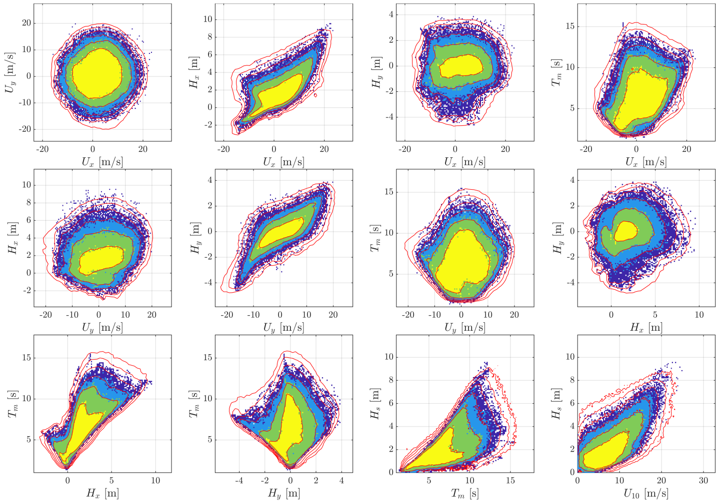
$$f_{R,\mathbf{W}}(r, \mathbf{w}) \approx f_{\mathbf{W}}(\mathbf{w}) \zeta f_{\mathrm{GP}}(r - u(\mathbf{w}); \tau(\mathbf{w}), \xi(\mathbf{w})), \quad r > u(\mathbf{w}). \tag{3}$$

With estimates of $(u(\mathbf{w}), \tau(\mathbf{w}), \xi(\mathbf{w}))$, simulation from $f_{R|\mathbf{W}}$ is simple.

In combination with a model for $f_{\mathbf{W}}$ (see, e.g., Wessel et al., 2025), we can simulate from $(R, \mathbf{W})$ by drawing from (3) with prob. ζ and, otherwise, from the corresponding *non-extreme* empirical distribution.

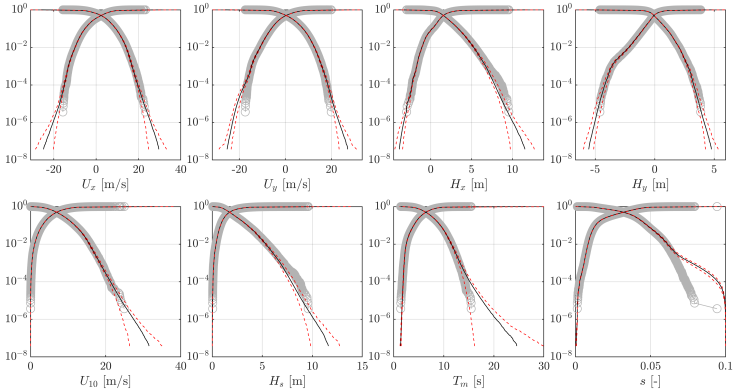
Wessel, Murphy-Barltrop, Simpson (2025). Generative machine learning for multivariate angular simulation. Extremes.

Estimates



Contours produced from deepSPAR model using 3000 years of reps.

Marginal diagnostics



Wind speed: $U_{10} = \sqrt{U_x^2 + U_y^2}$. Steepness: $s = 2\pi H_s/(gT_m^2)$.

Conclusion

- Deep regression models are conceptually simple yet powerful tools for modelling extremes - both **univariate** and **multivariate**.
- Despite their black box nature, neural networks can be used alongside statistical models to create something (atleast partially-)interpretable.
- Lots of code available on GitHub:
    - Jbrich95/pinnEV/USWildfireExtremes;
    - CASE2025_shortcourse/extQuantRegressDL/cde-RKeras-intro;
    - callumbarltrop/DeepSPAR/DeepGauge;
- Similar ideas can be used for modelling spatial extremes (see, e.g., Shao et al., 2025).

Shao, Richards, Huser (2025). Modeling nonstationary extremal dependence via deep spatial deformations. arXiv.

References I

Carreau, J. and Bengio, Y. (2007). A hybrid Pareto model for conditional density estimation of asymmetric fat-tail data. In *Artificial Intelligence and Statistics*, pages 51–58. PMLR.

Castro-Camilo, D., Huser, R., and Rue, H. (2022). Practical strategies for generalized extreme value-based regression models for extremes. *Environmetrics*, 33(6):e2742.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme Australian wildfires. *Spatial Statistics*, 59:100811.

Mackay, E. and Jonathan, P. (2023). Modelling multivariate extremes through angular-radial decomposition of the density function. *arXiv preprint arXiv:2310.12711*.

Mackay, E., Murphy-Barltrop, C. J. R., Richards, J., and Jonathan, P. (2025). Deep Learning Joint Extremes of Metocean Variables Using the SPAR Model. *Journal of Offshore Mechanics and Arctic Engineering*, 148(021201).

Majumder, R. and Richards, J. (2025). Semi-parametric bulk and tail regression using spline-based neural networks. *arXiv preprint arXiv:2504.19994*.

Papastathopoulos, I. and Tawn, J. A. (2013). Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143(1):131–143.

Pasche, O. C. and Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839.

Richards, J. and Huser, R. (2026a). Extreme Quantile Regression with Deep Learning. In de Carvalho, M., Huser, R., Naveau, P., and Reich, B. J., editors, *Handbook on Statistics of Extremes*. Chapman & Hall/CRC, Boca Raton, FL.

References II

Richards, J. and Huser, R. (2026b). Regression modelling of spatiotemporal extreme US wildfires via partially-interpretable neural networks. *JCGS, to appear*.

Richards, J., Huser, R., Bevacqua, E., and Zscheischler, J. (2023). Insights into the drivers and spatiotemporal trends of extreme mediterranean wildfires with statistical deep learning. *Artificial Intelligence for the Earth Systems*, 2(4):e220095.

Shao, X., Richards, J., and Huser, R. (2025). Modeling nonstationary extremal dependence via deep spatial deformations. *arXiv preprint arXiv:2505.12548*.

Short, K. C. (2017). Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Ed. Fort Collins, CO: Forest Service Research Data Archive.

Wessel, J. B., Murphy-Barltrop, C. J., and Simpson, E. S. (2025). Generative machine learning for multivariate angular simulation. *Extremes*, pages 1–49.

Fin.

SCAN ME

Scan for webpage.

PINN diagnostic

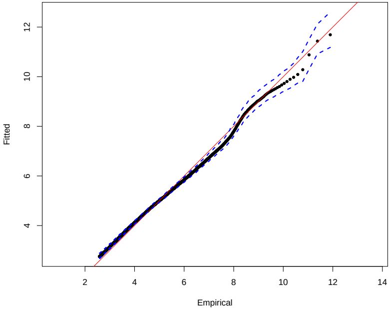


Figure: Pooled Q-Q plot for the local PINN bGEV-PP model fit on unit Exponential margins, averaged across all bootstrap samples. 95% tolerance bounds are given by the blue dashed lines.

PINN CV

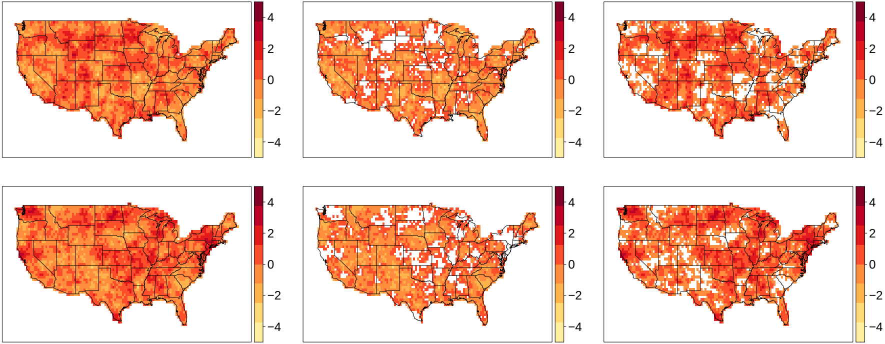


Figure: The left column provides two consecutive realisations $\{z(s,t) : s \in \mathcal{S}\}$ of the space-time Gaussian process. In the central column, spatial locations such that $z(s,t)$ is below the 0.2-quantile of $\{z(s,t) : s \in \mathcal{S}, t \in \mathcal{T}\}$ are assigned to the validation set; these are denoted by the white pixels. For the right column, white pixels denote spatial locations assigned to the testing set, where $z(s,t)$ exceeds the 0.8-quantile of $\{z(s,t) : s \in \mathcal{S}, t \in \mathcal{T}\}$.