Partially-interpretable Neural Networks for Extreme Quantile Regression

Jordan Richards and Raphaël Huser

CEMSE Division, King Abdullah University of Science and Technology (KAUST).

E-mail: jordan.richards@kaust.edu.sa

SCAN ME

Motivation

To build (extreme/quantile) regression models that balance:

- interpretability,
- scalability,
- high predictive accuracy,
- asymptotic and theoretical justification,

which can jointly estimate extreme quantiles and identify drivers of environmental risk.

Partially-interpretable Neural Networks (PINNs)

Let $\{Y(s,t) : s \in \mathcal{S}, t \in \mathcal{T}\}$ be a spatio-temporal process, and let $\mathbf{x}(s,t)$ be a d-dimensional vector of predictors observed at (s,t).

The PINN framework models some function or statistical parameter, denoted $\theta(s,t)$, of $Y(s,t)$ by splitting $\mathbf{x}(s,t)$ into two complementary subsets:

- "interpreted" predictors - $\mathbf{x}_{\mathcal{I}}(s,t)$,
- "non-interpreted" predictors - $\mathbf{x}_{\mathcal{N}}(s,t)$,

with the indices that map $\mathbf{x}(s,t)$ to these subsets consistent across all (s,t). Then we let

$$\theta(s,t) = h[m_{\mathcal{I}}\{\mathbf{x}_{\mathcal{I}}(s,t)\} + m_{\mathcal{N}}\{\mathbf{x}_{\mathcal{N}}(s,t)\}],$$

where h is a link function and

- $m_{\mathcal{I}}$ is a readily-interpretable function, e.g., linear or additive,
- $m_{\mathcal{N}}$ is assumed unknown and highly non-linear.

We estimate $m_{\mathcal{N}}$ using a neural network, e.g., densely-connected, convolutional, recurrent.

Models

We fit the following statistical models (implemented in the R package pinnEV):

- GPD$(\sigma, \xi > 0)$, with $\theta(s,t) := \sigma(s,t)$,
- bGEV$(\mu, \sigma, \xi > 0)$, with $\theta_1(s,t) := \mu(s,t)$ and $\theta_2(s,t) := \sigma(s,t)$
- (extreme value) point process model,
- single τ-quantile estimation, with $\Pr\{Y(s,t) < \theta(s,t)\} = \tau$,
- Bernoulli, when $Y(s,t) \in \{0,1\}$.

These are trained by minimising the negative log-likelihood associated with the above models.

Application to wildfire risk

We model the occurrence and extremes of wildfire spread in the U.S. and the Mediterranean.

- Response $Y(s,t)$ is monthly aggregated burnt area for a spatial grid-box,
- $d > 30$ predictors, including:
  - meteorological variables from ERA5,
  - land cover types from COPERNICUS,
  - orography, such as elevation and terrain roughness,
- We interpret the effect of vapour pressure deficit, 2m temperature and a drought index.

Related and further work

- GCNNs for irregularly spaced data
- Saudi Arabian $PM_{2.5}$ concentration modelling
- Implementation of new response distributions (suggestions welcome!)

Results

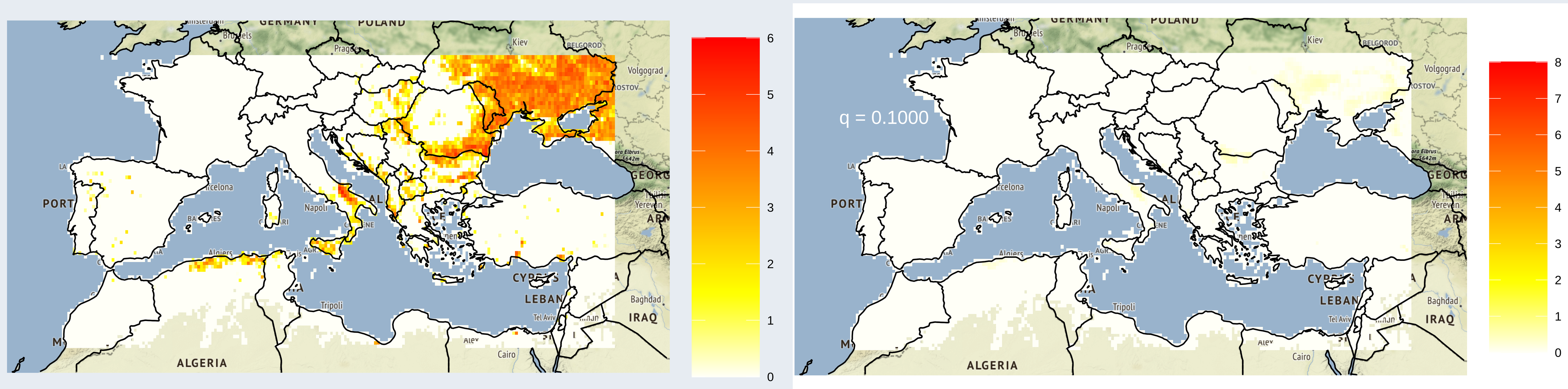

Figure: Monthly $\log(1+$ burnt area) for August 2008. Observation (left) and estimated extreme q-quantiles (right).

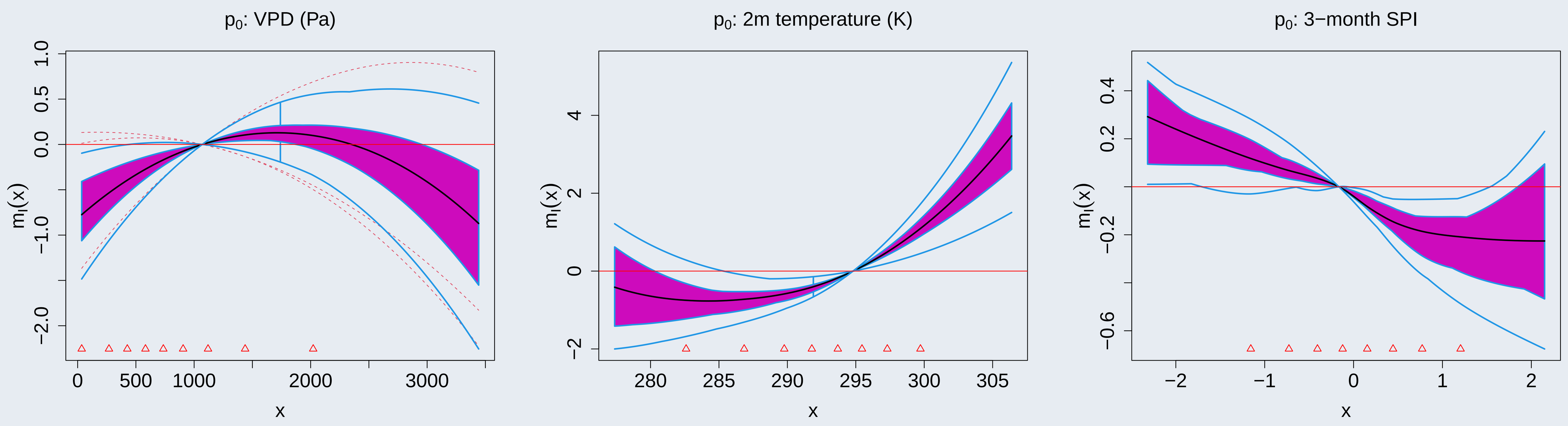

Figure: Functional box-plots for the impact of interpreted predictors on the log-odds of wildfire occurrence probability, p_0.

References

[1] J. Richards.
pinnEV: Partially-Interpretable Neural Networks for modelling of Extreme Values, 2022.
R package.

[2] J. Richards and R. Huser.
A unifying partially-interpretable framework for neural network-based extreme quantile regression, 2022.
arXiv:2208.07581.

[3] J. Richards, R. Huser, E. Bevacqua, and J. Zscheischler.
Insights into the drivers and spatio-temporal trends of extreme Mediterranean wildfires with statistical deep-learning, 2022.
Ongoing.