M2 Internship: Impact studies of extreme oceanic conditions on coastal risk indicators

Contacts:
Gloria Buriticá - gloria.buritica@agroparistech.fr

2025-2026

Keywords: Extreme value statistics
Application deadline: April 2026.
Context: This internship is part of a research project we are currently consolidating with Jérémy Rohmer (French Geological Survey, BRGM) et Anne Sabourin (Centre Borelli, ENS Paris-Saclay).

To apply, send a CV, a motivation letter and an academic transcript with the subject "Application for extreme impact studies internship".

1 Context

To anticipate the consequences of climate change, it is important to fully quantify impacts of extreme climate on society in terms of potential threats to people or infrastructures. This internship centers on developing machine learning (ML) prediction tools for conducting extreme impact studies. ML methods can learn complex relationships between the predictors and the response when enough data is available Hastie et al. (2005). Yet, such methods can't extrapolate to scenarios with limited data and are not reliable for prediction under extreme conditions Pasche et al. (2025). In the context of climate change, extreme climate conditions are expected to become more frequent and potentially threatening due to rising temperatures and increasing daily rainfall amounts on Climate Change (IPCC). Quantifying the impacts of an extreme climate variable X, say wind speed, on a target variable Y, say significant wave heights, is key for a complete risk analysis, but the most severe episodes happen rarely, which means there are only few historical observations to learn the relation between one extreme covariate X and the response Y. Classical ML regression methods will yield inaccurate predictions in this setting as predicting the distribution of Y given the predictor X when this last takes its highest levels becomes challenging due to data scarcity. Additionally, severe meteoceanic conditions are likely to produce a high-impact response. This relation of extremal dependence is well studied in the field of extreme value theory.

The purpose of the internship is twofold: deriving suitable theoretical limiting models of the impact of an extreme covariate on the response and also proposing statistical tools for enhancing prediction under extreme climate conditions. To address these goals we bring forward tools in extreme value statistics. The applicability of the field to tackle ML problems, particularly in regression, has only recently been studied and the few references focus either on quantile regression for critical probability levels (Gnecco et al. (2024), Chavez-Demoulin and Davison (2005), Castro-Camilo et al. (2022)), or mean-squared loss minimization Clémençon et al. (2025) without considering quantile regression. The recent work in Buriticá and Engelke (2024) studies domain generalization problems in ML for median point prediction. The internship aims at enhancing distributional regression methods in ML to better predict the response of extreme climate conditions.

2 Scientific description of the internship

We assume a statistical framework describing the relationship between a response variable Y and the vector of explaining variables $\mathbf{X}$. Our modeling assumption is that the generic vector $(Y, \mathbf{X})$ satisfies the

stochastic relation:

$$\mathbf{X} = \epsilon_{\mathbf{X}},$$
$$Y = g(\mathbf{X}, \epsilon_Y),$$

$\epsilon_{\mathbf{X}}$ (ϵ_Y) is a random vector describing the predictors (response) sampling process, $g : \mathbb{R}^{d+1} \to \mathbb{R}$ is a link function relating the predictors to the response and we access data: $((Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n))$ collected as identically distributed realisations of the given stochastic model. This means they have the distribution of the generic vector $(Y, \mathbf{X})$. We then translate our objective into a statistical task: learning the conditional quantile function of Y given $\mathbf{X}$, defined as:

$$Q_{Y|\mathbf{X}}(\tau) := \inf_{y \in \mathbb{R}} \{\mathbb{P}(Y \leq y \mid \mathbf{X}) \geq \tau\}$$

for $\tau \in (0, 1)$. This means building a predictor using the data at hand. Note the conditional distribution depends both on the link function g and also on the distribution of the data. Both of these functions can be arbitrarily complex and non-linear which makes estimation a difficult task without any a priori restrictions on the data-generating process.

We aim to study limiting predictive structures that enable the prediction of Y given that X is extreme. Our main focus will be on quantile regression of Y conditional to unusually large values of the norm of the covariate $\|\mathbf{X}\|$. Our working assumption Rootzén et al. (2018) will be a maximum domain of attraction condition. This assumption describes the extreme events of suitably standardize vectors in terms of an exponential random variable encoding the magnitude of the event, and a spectral vector describing suitable extremal directions. Note that standardizing margins to the Laplace scale is a common practice in extreme value theory. The maximum domain of attraction assumption is a mild and widely used assumption in extreme value analysis, which is equivalent to assuming a limiting joint distribution for affine rescaled componentwise maxima, known as the multivariate maximum domain of attraction condition. It is at the heart of extreme value theory De Haan and Ferreira (2006).

3 Scientific goals of the internship

The internship aims at enhancing predictions of a response under rare extreme conditions of the covariates. The first step consists in deriving suitable limiting predictive structures of the response Y under extreme covariates $\mathbf{X}$ relying on the maximum domain of attraction condition. The second step is to propose algorithms that integrate the information on the limiting predictive structures. Localising-based algorithms as nearest neighbors and random forests methods will be adapted. The methodology will be validated in simulations and applied to study the coastal impact due to extreme cyclonic conditions. The application's goal is to better predict the significant wave height of waves in Gaudeloupe (French Antilles) based on the physical characteristics of cyclones including wind speed and radius measures Rohmer et al. (2023). The data will be provided by BRGM.

During the internship the student is expected to perform the following tasks:

1. Conduct a literature review on the modeling assumptions on the extremal dependencies of the data from the extreme value statistics field;
2. Study the limiting predictive structures of the models from the domain of attraction assumption,
3. Implement ML methods that integrate the limiting predictive structures,
4. Conduct simulation studies to evaluate the methodologies and implement the methods on the case study of the impacts of extreme meteoceaninc conditions on coastal risks

The outcome of this internship is a study of the predictive structures and the implementation of simple localizing-based algorithms (e.g. nearest neighbors or random forests). A simulation study will be conducted to validate the results and the case study will be analysed.

4 Profile & environment

The candidate should be a 2nd year master or last year engineer student, in Mathematics or Statistics.

- Location : UMR MIA Paris-Saclay, Palaiseau Campus, 22 place de l'agronomie, 91120 Palaiseau.
- Supervision : Gloria Buritica is a researcher in extreme value theory with an expertise in climate applications and extrapolation regression algorithms.
- Starting date: flexible, starting in April 2026 or after.
- Duration: 6 months.
- Salary: as an intern, you'll receive a gratification which is capped around 650 euros/month net.

The candidate will have an office, and benefit from the work environment of the MIA Paris-Saclay laboratory, with many PhD students & postdocs working on statistical modeling and machine learning for the life and environmental sciences.

References

Gloria Buriticá and Sebastian Engelke. Progression: an extrapolation principle for regression. *arXiv preprint arXiv:2410.23246*, 2024.

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

Valérie Chavez-Demoulin and Anthony C Davison. Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(1):207–222, 2005.

Stephan Clémençon, Nathan Huet, and Anne Sabourin. On regression in extreme regions. *Electronic Journal of Statistics*, 19(2):4784–4828, 2025.

Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer, 2006.

Nicola Gnecco, Edossa Merga Terefe, and Sebastian Engelke. Extremal random forests. *Journal of the American Statistical Association*, 119(548):3059–3072, 2024.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Intergovernmental Panel on Climate Change (IPCC). *Technical Summary*, page 35–144. Cambridge University Press, 2023.

Olivier C Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1):e240033, 2025.

J. Rohmer, A. G. Filippini, and R Pedreros. ombining uncertain machine learning predictions and numerical simulation results for the extreme value analysis of cyclone-induced wave heights–application in guadeloupe. *Ocean Modelling*, 186:102275, 2023.

Holger Rootzén, Johan Segers, and Jennifer L Wadsworth. Multivariate generalized pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131, 2018.