PEPR EXPLORATOIRE
APPEL À PROJETS
2025

PEPR MATHS-VIVES

DOCUMENT PRÉSENTATION PROJET

GEO-ALERTX

Acronym / Acronyme	**GEO-ALERTX**		
Titre du projet en français	Machine Learning pour l'Alerte aux Risques Géophysiques		
Titre du projet en anglais	Machine learning risk-warning forecasts of geophysical natural hazards		
Keywords/Mots-clefs (min 5-max 10)	machine learning; extreme value statistics; multivariate regular variation; out-of-domain generalization; conformal inference; distributional regression; natural risks; early-risk warning systems		
Axes / Axes	☐ *Vivant* X Environment ☐ Society ☐ Emergence		
Leading institution /Établissement porteur	AgroParisTech		
Scientific coordinator / Responsable du projet	**Last Name, First name, Position / Prénom, Nom, Qualité**		
	BURITICA, Gloria, enseignante chercheuse		
	e-mail / Courriel	**Téléphone**	
	gloria.buritica@agroparistech.fr	+33 7 51 05 33 61	
Second scientific coordinator / Second responsable du projet	**Last Name, First name, Position / Prénom, Nom, Qualité**		
	ROHMER, Jérémy, chercheur		
	e-mail / Courriel	**Téléphone**	
	J.Rohmer@brgm.fr	+33 01 40 58 89 17	
Project duration / Durée du projet	60 Months		
Requested funding / Aide totale demandée	507.4 €	**Full cost / Coût complet**	507.4 €

Partner institution(s) involved in the project / Liste des établissements du consortium :

Name of the institutions of higher education and research / Établissements d'enseignement supérieur et de recherche	***Activity / Secteur(s) d'activité***

AgroParisTech, Université Paris-Saclay, MIA Paris-Saclay	Engineering school in agriculture and environmental sciences.
Ecole Normale Supérieure Paris-Saclay, Centre Borelli	Multidisciplinary school of higher education.

Name of the research organisations / Organismes de recherche	***Activity / Secteur(s) d'activité***
BRGM, French Geological Survey, EPIC	Institution for Earth Science and management of surface and subsurface resources and risks.

Research units involved in the project / Liste des unités de recherche impliquées :

Research units / Unités de recherche	**Établissement de tutelle**
MIA Paris Mathématiques et Informatique Appliquées - Paris-Saclay, RNSR 201119642G	AgroParisTech
Centre Borelli, RNSR 202023521J	Ecole Normale Supérieur Paris-Saclay
EPIC, RNSR 195922846R	BRGM, French Geological Survey

Institution / Liste des projets liés :

If applicable: Project links with existing France 2030/PIA entities (e.g. Labex, etc.) **Le cas échéant, listes des projets financés par le France 2030/PIA auxquels ce projet est lié (Labex, etc.)**	PEPR IRIMA - IRIMONT
If applicable: Project links with other existing projects **Le cas échéant, autres projets auxquels ce projet est lié**	ANR Mathématiques : EXSTA ANR Multi-thématique TANDEM ANR JCJC PROBTOOL

Résumé du projet en français

Le projet vise à fournir des méthodologies d'apprentissage automatique (machine learning - ML) open-source et fiables pour la prévision des risques naturels, assorties de garanties théoriques et exploitables à travers une grande diversité d'études de cas en Géosciences (phénomènes, évolution espace,temps). Malgré leur popularité, les méthodes de ML ne sont pas capables d'extrapoler à des scénarios où les données sont limitées et ne sont pas fiables pour des prédictions dans des conditions inédites. Dans le contexte des systèmes d'alerte précoce, les données historiques sont rares car les scénarios à risque surviennent rarement, et l'objectif principal est souvent d'anticiper des aléas sans précédent. Même lorsque des simulations numériques issues de modèles météorologiques et/ou géophysiques sont disponibles pour émuler ces conditions critiques, leur coût de calcul élevé limite également l'accès aux données.
Comme livrables principaux, nous fournissons des analyses de risques naturels critiques : les cyclones, les scénarios d'éboulements rocheux et les tsunamis induits par des séismes (**Task A**). Les études de cas documenteront le protocole de déploiement des méthodes par les utilisateurs, l'évaluation des modèles statistiques proposés, et fourniront des interprétations des prévisions d'alerte basées sur l'expertise des processus géologiques et météorologiques. Nous visons à évaluer la pertinence et l'expressivité des nouvelles méthodes de ML au regard de la dynamique géophysique. Les études de cas motivent également 4 questions scientifiques que nous abordons dans ce projet. Les questions **Q1-Q2** visent à améliorer les prédictions dans des situations complexes (grand nombre de variables, évolution spatio-temporelle). Les questions **Q3-Q4** visent à construire des indicateurs de confiance dans l'utilisation de ces techniques en fournissant des garanties et une quantification de la fiabilité des prédictions.
La seconde contribution de nos travaux s'adresse à la communauté de l'apprentissage statistique en établissant les garanties mathématiques sur lesquelles reposent les nouvelles méthodes de ML, car celles-ci seront conçues pour prédire des variables de réponse même lorsque peu d'observations extrêmes sont disponibles. Ce défi s'inscrit dans le cadre plus large de la généralisation du domaine d'apprentissage,un sujet qui suscite un intérêt croissant ces dernières années [68, 60, 11, 6]. Relever ce défi est crucial pour améliorer la fiabilité des systèmes d'alerte précoce basés sur le ML, en particulier lorsque des prédictions doivent être faites pour des événements extrêmes multivariés rares ou jamais observés auparavant. Un intérêt particulier sera accordé aux méthodes de régression prenant en entrée des indicateurs issus de variables extrêmes multivariées et temporelles/spatio-temporelles (**Task B→Q1-Q2**), ainsi qu'à la conception d'outils de validation et d'évaluation mesurant la fiabilité des prédictions (**Task C→Q3-Q4**).
Le projet contribuera à l'axe Environnement du PEPR Math-Vives à travers des méthodes statistiques fiables et guidées par les données (data-driven), ainsi que des logiciels reproductibles pouvant être utilisés pour prédire les impacts de scénarios extrêmes. Nous anticipons de fortes interactions avec les projets cibles **CLIMATHS**, **COMPLEXFLOWS** et **HYDRAUMATH**, qui visent à améliorer la modélisation mathématique et la simulation numérique des variables environnementales. Pour autant, notre projet est complémentaire des leurs, car il se concentre sur l'exploitation conjointe des modèles numériques et des enregistrements historiques afin de couvrir un plus large éventail d'applications, et ce indépendamment du fait que les variables cibles soient directement issues d'un modèle numérique ou non lorsque des données observationnelles sont disponibles.

Résumé du projet en anglais

The project focuses on delivering reliable open-source machine learning (ML) methodologies for predictions of natural risks, accompanied with theoretical guarantees, and exploitable across a diverse range of case studies in Geosciences (physical processes, evolution space-time). Despite their popularity, MLmethods can't extrapolate to scenarios with limited data and are not reliable for prediction from unseen conditions. In the context of early risk-warning systems, historical data is scarce because risky scenarios happen rarely and often the main goal is anticipating unprecedented hazards. Even when numerical simulations from weather and/or geophysical models are available to emulate critical conditions, they have a high computational cost which also limits access to data.

As primary deliverables, we provide in-depth analyses of critical natural risks: cyclones, rockfall scenarios, and earthquake-induced tsunamis (**Task A**). The case studies will document the users' protocol to deploy the methods, the assessment of the statistical models proposed, and will provide interpretations of the warnings' forecasts based on expertise from geological and weather processes. We aim to evaluate the novel ML methods' relevance and expressivity regarding geophysical dynamics. The case studies also motivate 4 scientific questions that we address in this project. **Q1-Q2** aim at improving the predictions in complex situations (high number of variables, space-time evolution). **Q3-Q4** aim at building trust indicators in the use of these techniques by providing guarantees, model tests and a quantification of the reliability in the prediction.

The second contribution of our work targets the statistical learning community establishing the mathematical guarantees on which the novel ML methods rely as they will be designed to predict response variables even when few extreme observations are available. This challenge falls within the broader framework of out-of-domain generalization, a topic which has received increasing interest in recent years [68, 60, 11, 6]. Addressing it is crucial for improving the reliability of early ML-based warning systems, particularly when predictions must be made about rare or previously unobserved multivariate extreme events. A particular interest will be given to regression methods taking input indicators from extreme multivariate and temporal/spatio-temporal variables (**Task B⇢Q1-Q2**), and to designing validation tools together with assessment tools measuring how reliable the predictions are (**Task C⇢Q3-Q4**).

The project will contribute to the Environmental axis from the PEPR Math-Vives through reliable data-driven statistical methods and reproducible software that can be used to predict the impacts of extreme scenarios. We anticipate strong interactions with the core projects **CLIMATHS**, **COMPLEXFLOWS** and **HYDRAUMATH**, which aim to improve mathematical modeling and numerical simulation of environmental variables. Yet, our project is complementary to theirs, as it centers on developing statistical tools that can later be coupled with both numerical models and historical records in risk analysis applications. This approach enables a broader range of applications, regardless of whether the target risk variables are derived directly from a numerical model, provided that some data is available on the link between predictors and the target

Contents

1 Context, objectives and previous achievements

1.1 Context, objectives and innovative features of the project

1.1.1 Context, Objectives and Multi-disciplinary Synergy

Despite significant efforts to reduce risks from natural disasters, global losses have escalated in recent decades [48, 27]. Water-related hazards accounted for 50% of all disasters between 1970 and 2019 [27], while major earthquakes continue to claim high death tolls, with five single-events causing over 50,000 fatalities since 2000. In this critical context, Early Risk-Warning Systems that forecast the impacts of extreme events are an essential rapidly evolving field [62]. To this end, Machine Learning (ML) approaches are becoming standard prediction tools [46]. Yet, despite the growing adoption of AI to predict responses of natural hazards, trust in its accuracy remains a major concern [36]. Formally, this questions statistical guarantees of the ML approaches, specifically their capacity to handle data scarcity and generalize to unseen, rare conditions, a topic which remains largely unexplored [54].

The main objective of this project is to develop reliable ML methodologies for forecasting of natural risks under extreme conditions, exploitable across a diverse range of natural hazards and implemented in open-source software. Our contributions bridge two complementary pillars:

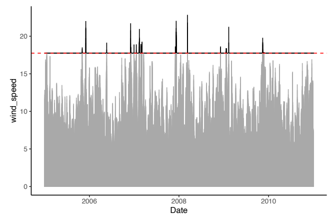
- **Pillar 1: Critical Natural Risks Evaluation (Task A).** We will provide detailed analyses of three primary natural hazards: cyclones, rockfall scenarios, and earthquake-induced tsunamis. These case studies will document ML user protocols, assess proposed statistical assumptions, and interpret warnings based on geophysical and weather processes.

- **Pillar 2: Mathematical Learning Foundations (Tasks B & C).** We will establish the mathematical guarantees for novel ML regression methods built to take input indicators from extreme multivariate and temporal/spatio-temporal variables (**Task B**), alongside validation tools and assessment metrics to measure forecast reliability (**Task C**).

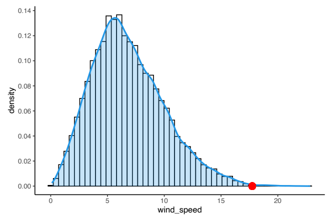
Our project contributes to the **Environmental** axis of PEPR Math-Vives offering a robust statistical framework for bias-correction of rare events' numerical simulations and prediction of critical risk indicators even when they are not outputs of geophysical models by leveraging observational records. Strong synergies are foreseen with core projects **CLIMATHS**, **COMPLEXFLOWS**, and **HYDRAUMATH**: GEO-ALERTX can be seen as their data-based counterpart. Furthermore, extremal regression algorithms can help anticipate climate change effects by extrapolating current impacts of extreme events towards future ones [48], creating logical bridges with **AGROSTAT**, PEPR TRACCS and PEPR IRIMA. Finally, our project resonates with recent theoretical advances in statistical learning and AI for extremes reviewed in [24, 16, 23]. Our focus will be on developing concrete operational tools for supervised extremal regression guided by practical questions. The methodological blocks include distribution-free uncertainty quantification, temporal dependencies, and model verification to facilitate users' accessibility of ML methods for extremal regression.

1.1.2 Risk Prediction Under Critical Conditions and Core Scientific Questions

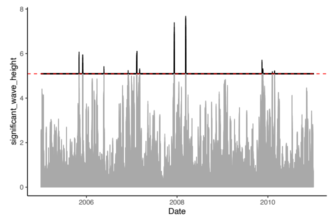
For numerous weather and geophysical applications, the extreme covariates have a high-impact on the response of risk indicators, see Figure 1 for a toy example relating wind and wave height illustrating the challenges of statistical regression on large covariates and data scarcity, which suggests an extremal dependence between predictors and the response (as modeled in (2)). Our main focus will be on a distributional forecast of a target Y conditional to unusually large values in the entries of the d-dimensional covariate vector $\mathbf{X} \in (\mathbb{R}^d, |\cdot|)$.



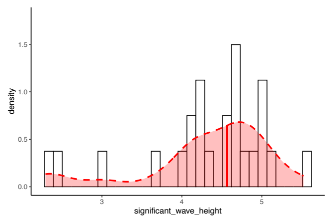
(a) X = wind intensity offshore



(b) x



(c) Y = wind wave heights



(d) Empirical conditional distribution of Y given x

Figure 1: Coastal hindcast data from Brittany [1] calibrated with observations. Left: Extreme records above the red line highlighted. Right: Histograms for wind intensity X (top) and empirical conditional distribution of wave heights Y given an extreme $X = 17m/s$ (bottom). The latter highlights the data scarcity issue in conditional estimation and shows that high wind intensity X drives large wind wave heights Y.

Building upon [13] and [6], a major focus of this project is to develop regression frameworks for extreme multivariate predictors and an unbounded target Y, in other words to construct prediction functions $q(x)$ aiming at minimizing the conditional risk as $u \to \infty$:

$$R_\infty(q) = \limsup_{u\to\infty} \mathbb{E}[\ell(Y, q(\mathbf{X})) \mid |\mathbf{X}| > u]. \tag{1}$$

With unbounded targets a natural loss function ℓ is the pinball loss ℓ_τ for a quantile level $\tau \in (0, 1)$ (see [40]). To model extremal dependence (cf. [61]), we assume the random vector $(Y, \mathbf{X})$ satisfies

$$\mathbb{P}\big((Y^* - u^*, \mathbf{X}^* - u^*) \in \cdot \mid \max\{Y^*, |\mathbf{X}^*|\} > u^*\big) \xrightarrow{d} \mathbb{P}\big(E + \mathbf{S} \in \cdot\big), \quad u^* \to \infty, \tag{2}$$

where $|\mathbf{X}^*| := \max\{X_1^*, \dots, X_d^*\}$, $Y^* = Q_E \circ F_Y(Y)$ and $X_j^* = Q_E \circ F_j(X_j)$, for $j = 1, \dots, d$, are standardized to the exponential scale. F_E, Q_E, are the exponential distribution and quantile functions, respectively, and E is an exponential random variables independent of the random vector $\mathbf{S}$ on $[-\infty, 0]^{d+1}$ satisfying $\max\{S_Y, S_1, \dots, S_d\} = 0$ a.s. Equation (2) is equivalent to assuming a limiting joint distribution for affine rescaled componentwise maxima, known as the multivariate maximum domain of attraction condition, at the heart of extreme value theory [19, 58].

Leveraging extreme value theory, we aim to provide robust predictions to support decision making for crisis and risk management regarding three phenomena that might potentially lead to huge economical or human impacts, namely: tropical cyclones, tsunamis and rockfalls. To guide the operational deployment of these novel ML approaches, the project addresses four central scientific and technical questions:

Q1) Variable Selection & Interpretability: How can we integrate dimension reduction techniques to identify influential predictors in extreme regimes (**Task B1**), and translate them into expert geophysical knowledge for risk analysis (**Task A**)?

Q2) Spatio-Temporal Dependencies: How can we account for complex temporal trajectories of extreme variables in particular weather-related ones that are crucial for cyclone tracking and impact forecasting (**Task B2**, **Task A**)?

Q3) Hypothesis Testing & Validation: How can we design statistical tests to help practitioners verify the validity of the assumptions made, notably if the asymptotic multivariate maximum domain of attraction condition holds true on finite-sample datasets before extrapolation (**Task C1**, **Task A**)?

Q4) Uncertainty Quantification: How can we provide actionable reliability metrics for generated alerts using modern statistical frameworks like conformal prediction [26, 2] (**Task C2**, **Task A**)?

By constructing a structured workflow between AgroParisTech, BRGM, and Centre Borelli, this project guarantees the developed open-source softwares and statistical tools are directly in line with the constraints and needs for crisis and risk management in the domain of natural disasters.

1.2 Main previous achievements

ML methodologies for prediction of high impact events BRGM has undertaken different statistical learning approaches for the prediction of cyclone-induced waves in the French West Indies [59] from the characteristics of the cyclones, and for the predictions of tsunami-induced coastal impacts in the Mediterranean sea [60] from the characteristics of the earthquake, and for prediction of deposit area induced by rockfalls (ongoing work within PEPR IRIMA IRIMONT) from the rockfall source characteristics. These studies however have not addressed the core problem of predictions of impacts of unprecedented extremes as well as the reliability of ML predictions, which has arisen in the exchanges with end-users of early-warning systems and end-users of probabilistic hazard maps related to these phenomena.

Regression on extreme covariates This research area has been highly active, with key foundational contributions from the project leaders. A. Sabourin and co-authors established a rigorous statistical learning framework to derive the guarantees on (1), generalizing from binary classification [39, 14] to least-squares [13] and Lasso regression [16]. While these works laid essential foundations, they focus on mean regression, which is ill-suited for unbounded targets with infinite moments. Recent extensions introduced quantile regression via SVMs [43] to handle unbounded targets and regularization strategies; however, this framework lacks variable selection, limiting its operational interpretability. Furthermore, all these models assume full asymptotic dependence, requiring all covariates to become large simultaneously, which is unlikely in high-dimensional settings. To address this, integrating preliminary dimension reduction is a core objective of **Task B**. Unsupervised dimension reduction for multivariate extremes is a major focus in EVT [24], where A. Sabourin and co-authors have pioneered key advancements (e.g., [35, 21]). Multivariate time series modeling will also play an important role, G. Buritica and co-authors have contributed to this field [9, 8]. It will allow is to guide dimension reduction for temporal/spatio-temporal covariates.

In a separate direction and without constraints of asymptotic dependence of extremes, G. Buritica and co-authors [6] introduced a median regression framework for a univariate extreme predictor. They proved that the conditional median can be approximated via a first-order extrapolation function on the exponential scale, derived from a linear stochastic approximation. This structural approximation remains valid for a large non-parametric class of models, and its relative approximation error vanishes asymptotically, justifying restricting candidate predictive functions to this limiting structure.

Our project extends these univariate findings to multivariate, high-dimensional predictors $\mathbf{X}$ to identify structural constraints learnable from data. Demonstrating this feasibility, an ongoing master's internship (since May 2026 at AgroParisTech) is currently adapting ML algorithms to embed these limiting constraints.

2 Detailed project description

2.1 Project outline, scientific strategy

ML methods struggle to produce accurate predictions outside their training range. Traditional distribution-robust [47, 55] or causal-based methods [63, 11, 47] for out-of-domain generalization seldom leverage extreme value theory (EVT). While EVT historically focused on inference [17, 37, 18, 69], this project bridges EVT and ML to optimize prediction under extreme covariates. Incorporating EVT-grounded models into ML methods restricts regression functions in extreme regions allowing us to derive statistical guarantees.

One major challenge for addressing the applications that we envision, is the features given as an input can be numerous and have redundant information, but the effective data-sets are already small when we focus on extremes, which points to the curse of dimensionality, a well-known phenomenon from high-dimensional statistics [31]. In multivariate extremes, it is also widely established that 'typical' extremes in a data-set can be vectors where only subsets of coordinates are large simultaneously. The suitable subsets of extreme coordinates are known as 'directions', and in practice only a few of them are realistic, which points to the notion of sparseness [35, 34, 49, 50, 24]. Although unsupervised EVT extensively explores sparsity [35, 24], its application to supervised settings remains a major gap.

In the supervised setting, two problems can be of interest: either considering prediction of an extreme response Y, or consider a typical response value Y but under extreme covariates $\mathbf{X}$. For the first case, work in extreme value theory focuses on quantile regression [32, 8], including dimension reduction [28, 30]. Conversely, predicting a typical response Y given extreme covariates $\mathbf{X}$ is remains under-explored [13, 6], and associated dimension reduction frameworks are limited [16, 43].

From an application perspective, the major challenge is highlighting the benefits of appropriate ML methods for predicting impacts of extreme events, and investigating how to adopt ML tools in operational protocols for early-warning systems and probabilistic hazard assessment (**Task A**). From a theoretical perspective, our project addresses two major statistical bottlenecks: 1) Exploiting sparse extremal dependence structures to achieve dimension reduction under extreme covariates, while integrating spatio-temporal dependencies for distributional regression (**Task B**$\rightsquigarrow$**Q1-Q2**). 2) Designing statistical validation tools to assess model assumptions as well reliability indicators under extreme regimes (**Task C**$\rightsquigarrow$**Q3-Q4**).

Scientific Strategy The project is structured into three interconnected tasks running in parallel. **Task A** handles the operational case studies and triggers the methodological needs which are addressed through the theoretical developments of **Task B** and **Task C**, and then validated through the case studies. Strong interactions between group leaders will ensure a unified workflow across all axes.

Task A) Operational Early Risk-Warning ML-based Methods and Case Studies

A1) Protocol for performance analysis of real case predictions

A2) Transfer of methods and protocols to operational studies

Task B) Extremal Regression: Dimension Reduction and Temporal/spatio-temporal Covariates

B1) Algorithms for dimension reduction and their theoretical analysis

B2) Extensions to temporal/spatio-temporal covariates and time series prediction

Task C) Reliability Tools: Prediction Models and Forecasts in Extreme Regime

C1) Data-based assessment of the validity of working assumptions

C2) Reliable prediction intervals via conformal inference on extreme covariates.

2.2 Scientific and technical description

2.2.1 Task A) Operational Early Risk-Warning ML-based Methods and Case Studies

The objective of **Task A** is twofold. **Task A1** evaluates the predictive performance of the new ML extremal regression developments from **Tasks B, C** on three case studies related to natural risks. **Task A2** studies the viability of integrating the methods in operational studies at BRGM.

Three types of natural phenomena are tackled, covering a wide range of situations and diffent levels of complexity: cyclones: weather-related hazards (Case 1), rockfalls: geological-related hazards (Case 2), and tsunamis: geological-related hazards (Case 3). We define below the data, computer experiments and protocol for performance analysis.

Case 1 (cyclones): We focus on extreme waves induced by tropical cyclones [59]. We envision to rely on a database of synthetic cyclones that enables to relate the cyclone spatio-temporal evolution (cyclone eye space-time position, radius, atmospheric pressure, wind speed) and the induced sea states (significant wave height, peak wave period, peak direction) around the Guadeloupe archipelago in the French West Indies. This case is based on numerical simulation and ML modeling of the Interreg-CaribCoast project finalised in 2022 [3]. Case 1 has a high number of predictors (>30) and aims at predicting temporal/spatio-temporal evolution. This case will be studied through the developments of **Task B2**.

Case 2 (rockfalls): We focus on predicting maximum spatial extent of rock deposits after an occurrence of rockfalls. Two real applications will be considered, the case of Cassé de la Rivière de l'Est at La Reunion island from 2023 and 2024, and the Etaches event in the Alps in 2020; see [56]. The goal is to predict the spatial evolution of rock deposits depending on the initial rock volume and on the rheological parameters of the rock formations. This case benefits from the scientific expertise of the ANR JCJC project PROBTOOLS by M. Peruzzetto launched in May 2026 as well as results from the PEPR IRIMA IRIMONT. Case 2 has a low-to-moderate number of predictors (<10), and aims at predicting spatial evolution (map of deposit height after rockfall). This case will be conducted together with the developments of **Task B1**.

Case 3 (tsunamis): We study impacts of an earthquake-induced tsunami on the coasts. We use the database of [60] of the 1887 Ligurian tsunami to relate the earthquake source parameters that describe the fault rupture during the earthquake and the sea surface elevation at the coasts. This case benefits from the results of the AMI ANR TANDEM in 2019. It has a moderate number of predictors (around 10) and aims at predicting scalar variables of interest (maximum value of sea surface elevation). This case will serve as a benchmark dataset common to all **Tasks B and C**.

The primary goal of **task A1** is to assess if the new ML developments improve the prediction performance (prediction error and uncertainty) of impacts of extreme realistic scenarios. This means investigating to which extent the moist influential predictors selected by the methodology from (**Tasks B1, B2**) are informative about the geological and weather processes. In a second **task A2**, we aim at going beyond the performance analysis of **Task A1** by assessing the feasibility of integrating the new ML developments in BRGM operational studies, i.e., within early-warning systems and within probabilistic analysis used for the regulatory risk documents. This means considering practical criteria like implementation effort, computation time cost, assessment tools for validation protocols, and the level of expertise to use the new ML developments, etc. The assessment tools from **Task C** should help the end-users build trust in the use of these new ML developments in the decision-making process for crisis and risk management (e.g. [38]).

Task A1-A2: *Evaluate the performance of the ML developments by covering multiple geophysical and high-dimensional situations (***Task A1***).* **Task A2** *will investigate transferability to operational studies.*
Main investigators: J. Rohmer (Task leader), A. Lemoine, R. Pedreros, M. Peruzzetto, G. Buritica, A. Sabourin
Related publications from project members: *[60], [59], [56], [42], [3]*

2.2.2 Task B1) Extremal Regression: Algorithms for dimension reduction

The goal of **Task B1** is to contribute to question Q1 in 1.1.2 by providing guarantees on the new methods aiming to improve extremal predictions while identifying influential predictors through variable selection strategies. We apply the methods to cyclones, rockfalls and earthquake-induced tsunamis.

We explore a regression approach allowing automatic selection of the most relevant features (and in our case extremal features)—while encouraging sparsity or simplicity through penalty terms—which tends to be more effective [65]. A natural starting point is linear quantile regression with an ℓ_1 penalty trained on the $k << n$ (where n is the sample size) largest observations. This approach is motivated by assumption (2), which entails that in extreme regimes the variables behave according to the model $E + (S_Y, S_1, \ldots, S_d)$ which justifies a linear relation, and introduces sparsity through assumptions on the support of the spectral component $\mathbf{S}$. Figure 2 illustrates an example featuring only two extremal directions, this sparsity feature is common for heavy-tailed additive models [51]. Then, assuming only a few, say K, extremal directions are possible, the quantile regression function simplifies to a mixture of K linear models:

$$\hat{Q}_{Y^*|\mathbf{X}^*=\mathbf{x}^*}(\tau) \approx \sum_{k=1}^{K} \mathbf{1}\{\mathbf{S} = \mathbf{s}(k)\}\big(\beta^T(k)\mathbf{x}^* + \mu(k;\tau)\big), \quad ||\mathbf{x}^*| > u^*, \tag{3}$$

as $u* \to \infty$, where $\beta(k) \in \mathbb{R}^d, \mu(k;\tau) \in \mathbb{R}, \mathbf{s}(k) \in [-\infty, 0]^{d+1}$ and $\sum_{j=1}^{d} \beta(k) = 1$, for all $k = 1, \ldots, K$, for $\tau \in (0,1)$. We investigate a localizing-based technique to a) partition the predictors' space relative into neighbors defined as points with the same extremal direction and b) rely on a penalized lasso for quantile regression for prediction [10]. Extensions to conditional probability forecasting can also be investigated replacing the loss by the continuous rank propensity score CRPS; see [57, 70, 20]. Even though the initial focus of the project is quantile regression, probability forecasting opens the road for future investigation.

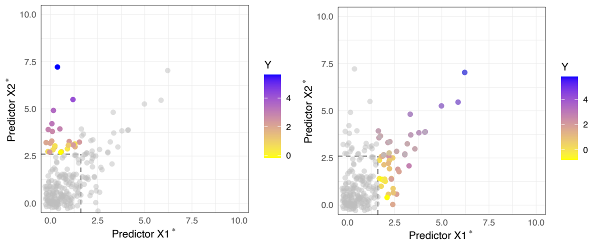


Figure 2: Scatter plot of predictors $\mathbf{X} = (X_1^*, X_2^*)$ from an additive model $(Y, \mathbf{X})$ with equivalent heavy-tails with $X_j^* = Q_E \circ F_j(X_j)$, for $j = 1, 2$. Each column highlights one of the two extremal directions with the magnitude of the response Y colored for the most extreme records. For f_1, f_2 polynomials, we plot $Y = f_1(X_1) + f_2(X_2) + \varepsilon_Y$.

Establishing theoretical guarantees of the methodology is the second main objective of this task. From a theoretical point of view, the main bottleneck is to show that the resulting prediction bounds align with classical ones, except that the total sample size must be replaced by an "effective" size k. The key technical challenge is that selecting the largest observations introduces dependence within the extreme sample, which we plan to overcome by leveraging concentration inequalities and techniques of proofs [44] suitable for low probability classes and learning problems based on the k largest observations.

Task B1): *Derive theoretical guarantees of localizing-based ML-methods detecting relevant extremal drivers and implement the ML methods in two case studies: Case 1 on cyclones with large number of covariates (>30), and Case 2 with moderate number of covariates (around 10).*
Main investigators: G. Buritica (Task leader), A. Sabourin, PhD hiring 1
Related publications from project members: *[43], [6], [13].*

2.2.3 Task B2): Extensions to temporal/spatio-temporal covariates

The objective of **Task B2** is to contribute to question Q2 in 1.1.2. We propose methodology to enhance extremal forecasting and prediction of cyclones (case 1), by integrating information on the temporal/spatio-temporal dependencies of predictors from trajectories of weather variables.

Covariates of cyclones are given as the temporal/spatio-temporal evolution of main physical predictors. The spatial resolution of cyclonic conditions is $> 10,000$ pixels and time steps are given every 3 hours and depend on the cyclonic's activity duration (between half a day and 10 years). To model temporal/spatio-temporal extremes, we propose to rely on the notion of maximum domain of attraction of time series and replace condition (2) with a the assumption of regular variation of time series; see [7, 4, 41, 51].

The first sub-task focuses on forecasting future values based on past extremal records. Existing tools focus on estimating the 'extremal index' (the average number of successive excedances) or conditional prediction [33] based on knowledge of one exceedance. However, predicting a future variable $\mathbf{X}(t+h)$ from a past window $(\mathbf{X}(t-J), \ldots, \mathbf{X}(t))$ during crisis scenarios—conditioned on $\mathbf{X}(t)$ exceeding a high threshold u—remains unaddressed within a non-asymptotic statistical learning framework. The major bottleneck lies in the breakdown of the independent and identically distributed (i.i.d.) assumption for training data. To resolve this, we will formalize prediction of extremes by developing algorithms backed by finite-sample generalization bounds. To overcome temporal dependence, we will operate under weak dependence (mixing) assumptions, where guarantees exist outside of extremes via stability or Rademacher complexities [52, 53]. A complementary approach will exploit the asymptotic independence between extreme clusters under "anti-clustering" conditions to adapt the "many short trajectories" paradigm [66].

The second sub-tasks studies optimal predictors. Drawing on [13, 7] and **Task B1**, we force methods to identify 'typical' temporal/spatio-temporal extreme episodes [7, 9] to induce dimension reduction strategies.

> **Task B2):** *Forecast time series after an extreme record, motivated by the case study of cyclones (Case 1) where the trajectories of weather variables can be informative about the expected coastal wave heights.*
> **Main investigators: A. Sabourin, G. Buritica (Task leaders), PhD hiring 1, envisioned collaboration with Cambyse Pakzad (McF, U. Paris Nanterre)**
> **Related publications from project members:** *[9],[7], [13]*

2.2.4 Task C1) Assessing the Validity of Working Assumptions

> The objective of **Task C1** is to contribute to Q3 in 1.1.2. This task develops a statistical test framework for validating the modeling assumption (2). The learning guarantees of our ML extremal regression methods rely on those, which need validation in the examples of (**Task A**) for an operational pipeline.

The primary objective of this task is to develop statistical tests to assess the validity of the limit model (2) and to guide threshold selection. This is particularly challenging in high dimensions, where the limit distribution of excesses—determined by the spectral variable S—may become degenerate and concentrate on subspaces of the domain. Existing literature such as distance covariance tests [71] provide limited statistical guarantees, while methods proposed in [22] are restricted to moderate dimensions. This research explores new avenues allowing for larger dimensions of the covariate. Specifically, we will investigate permutation tests [5] to evaluate the independence between the exponential and spectral components, alongside approaches based on classification and/or ranking procedures and ROC curve analysis [15, 45] to compare original and permuted sample distributions. Anticipated bottlenecks include the additional dependence introduced by restricting training sets to the $k \ll n$ pairs with the largest covariate norms, as well as a bias term arising from the non-asymptotic nature of observations that must be accounted for. Furthermore, since sparsity structures are expected at extreme levels [35], the common assumption of density existence [5] must be amended to accommodate densities restricted to subspaces of the ambient space. Crucially, in a supervised setting, the main goal is not to test the multivariate regular variation assumption itself, but rather to verify the validity of the approximate regression model [13]. Because this represents a weaker condition than (2), it may lead to simplified testing procedures with enhanced statistical power. The methods and theory developed in this task are generic and are meant to be applied to any real world datasets related to Geosciences, in collaboration with BRGM and F. Dias. Our collaboration will focus on applying multivariate extreme model validity tests to classic sea-state descriptors—such as spectral moments, steepness, skewness, and directional spreading. The goal is to determine if these parameters can be effectively used as extreme-valued covariates. Ultimately, this framework aims to predict rogue wave occurrences [25] or, at least, issue reliable warnings when the conditional probability of such events exceeds a safety threshold.

Task C1): *Develop statistical tests to reject scenarios where the multivariate regular variation assumption is not satisfied.*
Main investigators: A. Sabourin (Task Leader), PhD hiring 2, J. Rohmer, F. Dias. Envisioned collaboration with Stephan Clémençon (Pr, Télécom Paris)
Related publications from project members: [35, 13, 25]

2.2.5 Task C2): Constructing Reliable Prediction Intervals via Conformal Inference on Extreme Covariates

The objective of **Task C2** is to contribute to question Q4 in Section 1.1.2. It provides tools to assess the uncertainty of the generated forecast. Indeed, establishing an operational, uncertainty quantification of early risk-warning strategies is essential for building user trust.

In risk management, the relevance of AI models heavily relies on their capacity to quantify uncertainty through realistic prediction intervals. While conformal prediction [68, 67] addresses this need in a distribution-free framework, extending it to extreme extrapolation reveals two major bottlenecks: the loss of coverage properties due to the lack of representative data in the tails, and the necessity of achieving conditional coverage guarantees (local to the extreme region) whereas classical guarantees only hold marginally (on average across the entire space). To overcome these limitations, this project will develop robust conformal inference methods tailored to extreme regimes. Moving beyond the marginal framework, we will leverage recent advances in conditional conformal inference [29]. The core challenge will involve reintroducing dependence on covariates through residual regression methods specifically adapted to extremes. The expected outcomes will enable the integration of reliable prediction uncertainty quantification into critical environmental variable regimes for natural risk management. This question is critical across all three real-world use cases provided by BRGM, where end users and risk planners increasingly demand rigorous uncertainty quantification to assess prediction reliability. Collaborating closely with F. Dias will facilitate a wider dissemination of our findings and drive diverse applications across the geosciences, as a possible alternative to Bayesian approaches to assess uncertainty in extreme regimes [12]

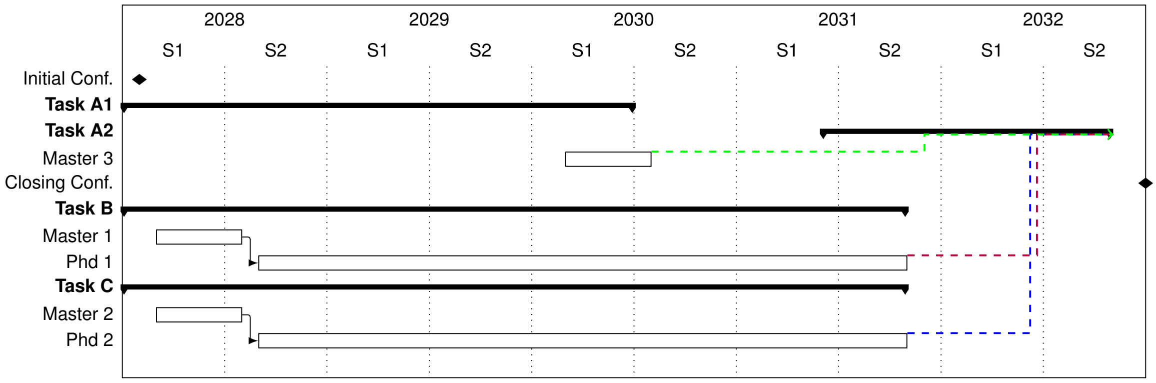
Task C2): *Construct reliable prediction intervals relying on Conformal Inference.*
Main investigators: Anne Sabourin (Task Leader), PhD hiring 2, J. Rohmer, F. Dias
Related publications from project members: [39, 13]

2.2.6 Scientific stakes

The Environnemental PEPR Math-Vives axis is centered around target projects aiming at improving modeling and simulation of geophysical variables. **CLIMATHS** project improves models coupling the ocean and the atmosphere, **COMPLEXFLOWS** improves mathematical modeling of complex natural flows as rockfalls and landsliding, **HYDRAUMATH** improves the coastal systems coupling water flux and land, **MATHSOUT** improves numerical simulation tools for underground flows. We foresee three main interactions the latter projects. First, when the risk indicator is a direct output, our methodology allows for statistical evaluation of responses under extreme covariates based on historical data. Second, when the indicator is not a direct output, our methods use the prediction function trained on historical data to extrapolate the impacts of future extreme scenarios on the response This interests PEPR **IRIMA** for landslides and rockfalls **IRIMONT**, and for cyclones and tsunamis (**IRICOT** and Outre-Mer). Moreover, in a climate change context the climate events that are considered as 'extreme' today might become more frequent in the future [48], then predicting their impacts requires regression tools that are reliable on extreme regions. Third, within projects analyzing climate change impacts, our methods can evaluate how future variations in extreme covariates affect the target variable, for example, **AGROSTAT** and **CLIMATHS** could benefit from our extremal regression tools. Our project can also contribute indirectly with the PEPR **TRACCS** (Extending)

2.3 Planning, KPI and milestones

2.3.1 Work plan:



A launching workshop is planned on February 2028 aiming at bringing together researchers in extreme value statistics and natural risks. A closing workshop will take place in October 2032 allowing the team members in the overall process to share their contributions with risk practitioners and identify the remaining challenges.

To foster interdisciplinary collaborations, the project will hire two PhD students. One will bring a background in statistics or geophysical sciences with excellent programming skills to contribute to **subtasks in A and B**. The second student will leverage a strong mathematical background to conduct the theoretical research in **Task C**. BRGM will work on the performance protocol of the ML developments **Task A1**. In the meantime, the students will benefit from expertise of BRGM on the three test cases with support on data processing, on the implementation/comparison of 'classical' ML models used in past studies, and support for conducting new numerical models via BRGM's high performance computing hub. Based on the results, BRGM will focus on the feasibility assessment and transfer to operational studies **Task A2**; the lessons drawn from this analysis will be presented at the closing workshop and exchanged with risk practitioners.

The recruitment of two PhD will be done upon completion of their M2 internships (March-July 2028). The first inter (AgroParisTech) will take in hand the preliminary data analysis of cyclones (*Case 1*). The second inter (Centre Borelli) will carry on **Task C**. One third M2 intern (AgroParisTech) will be hired on March-July 2030 to work on the earthquake-induced tsunamis (*case 2*). We detail below the M2 internships that will kick off **Tasks B and C** of the project and the data analysis of real cases 1-3 from **Task A**.

PhD hiring 1: *First, during an M2 she/he will analyse temporal and spatio-temporal extreme behavior of cyclones* **Task A1** *through adapted summary statistics [9], apply clustering algorithms to identify 'typical' extreme events, and adapt classical forecasting algorithms to account for extremal covariates to launch* **Task B**.
Main investigators: *G. Buritica (Task leader), M2 hiring 1 (⇝***PhD hiring 1***), A. Sabourin, J. Rohmer*

PhD hiring 2: *First, during an M2 she/he will kick off* ***Task C1*** *and explore the feasibility of permutation tests and ROC curve analyses for assessing the validity of multivariate EVT assumptions, with a particular focus on the use cases described in* ***Task A1****. This will pave the ground for the first PhD project, during which the construction of prediction intervals via conformal inference* ***(Task C2)*** *will subsequently be addressed.*
Main investigators: *A. Sabourin (Task leader), M2 hiring 2 (⇝***PhD hiring 2***). Collaboration with S. Clémençon (Télécom Paris) and use case analysis with J. Rohmer.*

The **third M2 hiring** *will develop ML methods for predicting rockfall hazard maps from rheological parameters and initial rock volume (Case 2,* **Task A1***). The response here is a high-dimensional spatial field, requiring models that explicitly account for spatial dependence in the outputs. Several classes of algorithms will be investigated, including convolutional neural networks (CNNs), U-Net architectures for spatial prediction, and more recent transformer-based or graph neural network approaches depending on the spatial representation of the data [64]. A key methodological objective will be to adapt these models when covariates are extreme.*
Main investigators: *G. Buritica (Task leader), M2 hiring 3, T. Rebafka, J. Rohmer*

2.3.2 Deliverables

The outcomes of the project can be quantified in terms of peer-reviewed research articles in high-standard journals in statistics and ML on one side, and methodological communications detailing the user protocols of the ML novel tools on three case studies for publication on weather and geophysical journals. The proposed ML methods for distributional regression will be compiled in an **open-source** well-documented and operational code available in R and Python using a Github repository of the project. Participation in conferences in natural risks and statistical learning are envisioned. Midterms and final reports together with communications for the Math-Vives target project members are also planned. The project also ensures the knowledge **transfer** towards younger generations, sharing expertise both in field domains and applied statistics via the closing workshop and the exchanges with the other PEPR projects (IRIMA, Maths-Vives).

2.3.3 Risks and pitfalls:

One main pitfall of ML research on out-of-domain generalization is the lack of benchmark data-sets. Existing methods' performance depend on the type of domain shifts: mean shifts, variance shifts, design shifts, interventions shifts, etc. We completely overcome this pitfall since our project relies on real case studies where the distributional shifts are well identified and correspond to extremal predictors. BRGM representatives in the project guarantees an expert's advice on the data-sets and risk from natural disasters. Current operational studies at BRGM on early-warning systems and probabilistic analysis of natural risks will ensure that the developments remain realistic with respect to the operational constraints and end-users' needs. A second pitfall we identified is the need of both engineering profiles capable of scaling the methodological developments, and profiles with strong theoretical background in statistics. For this reason, we accommodate the theoretical and applied tasks to be conducted by two different PhD hirings with the respective skills. An overlap of the two students will assure an exchange of knowledge between them and between the supporting members of the project. The proximity between AgroParisTech and the Centre Borelli where both students will be hosted will also enhance interactions between the applied and theoretical contributions of the project. The location of BRGM at Orleans, 1 hour 30 from Paris also facilitates exchanges. A final risk we have identified lies are the theoretical challenges arising from the scarcity of statistical learning research in extreme value analysis, specifically regarding testing model assumptions and producing prediction intervals with non-asymptotic guarantees.To mitigate this risk we count on key experts in extreme value theory and extremal statistical learning to unlock the theoretical obstacles.

3 Project organisation and management

3.1 Project manager

G. Buritica (AgroParisTech) coordinates the project. J. Rohmer (BRGM) leads the interface to applications.

Gloria Buriticá: She is an associate professor (MCF) at MIA-Paris Saclay, AgroParisTech since Sept. 2024, specializing on EVT and modeling temporal and multivariate dependencies for environmental applications. Her work covers out-of-domain generalization, heavy-tailed time series and risk assessment. Before this she held a Postdoctoral position at the University of Geneva (Oct 2022-Sept 2024). Previous to this, she defended her PhD in statistics at Sorbonne Universite - LPSM (Oct 2019-Sept 2022). The GEO-ALERTX project would be a starting point in her research career allowing her to establish and consolidate long-term collaborations with Borelli, BRGM, in view of future projects (ERC Starting or Consolidator). **Responsibilities:** Internship supervision (4 successful master internships, 2 ongoing), Workshop organization 'BIRS 26w5651' Banff, Canada, to come in Aug. 2026, 'Causality in Extremes', University of Geneva 2024. **Referee in:** Annals of Statistics, Bernoulli, Extremes, Journal of Applied Probability, Journal of Machine Learning Research. **Publications:** 5 articles in peer-reviewed journals, 1 submitted article and 2 R-package in GitHub. **Communications:** >10 Conferences, >10 seminar participation. (H-index of 5). CV in Sec 7.

Jeremy Rohmer is a research engineer at BRGM since 2005. His main research activities focus on uncertainty quantification in Earth Sciences. He defended his PhD thesis in 2015 on uncertainty quantification for natural hazard assessments. He has been scientific and/or work package coordinator of several studies at the interface between applied sciences and statistics (ANR RISCOPE, FP7 MEDSUV, FP7 ULTIMATE, H2020 NARSIS), and is currently ANR HOUSES project coordinator which aims to define a harmonized framework to reflect uncertainties along the modelling chain of Geodata. He is author and co-author of 80 peer-reviewed publications since 2009 (H-index of 16).

3.2 Organization of the partnership / Organisation du partenariat

The GEO-ALERTX consortium relies on a task-led structure designed to combine statistical learning with key natural risk applications. G. Buritica (Lead, **Task B**) brings expertise through her work on EVT applied to temporal and multivariate dependencies for environmental applications. G. Buritica's team is joined at AgroParisTech by T. Rebafka expert in ML and deep learning methods, network analysis and random graph models. She scales the project's framework to high-dimensional spatial responses (real case 2).

J. Rohmer (Lead, **Task A**) jointly with BRGM specific hazard experts (see table) complements this framework delivering advanced expertise in natural hazard risk assessment by integrating geophysical modeling with data-driven statistical methods.

A. Sabourin (Lead, **Task C**) is an expert of high-dimensional statistics and statistical machine learning for extreme value analysis. Her track record includes 30 peer-reviewed publications (H-index: 18) and her appointment as Chair of the Scientific Committee for EVA 2027 (Montréal). She also brings experience in project management as PI of the ANR PRC project EXSTA (2023–2027), which concludes before the start of GEO-ALERTX, ensuring full operational availability. A. Sabourin's team is joined at Centre Borelli by F. Dias (Pr, ENS Paris-Saclay, H-index: 70), an expert in fluid mechanics and ocean engineering whose research has been supported by four European Research Council (ERC) Grants (including MULTIWAVE (2011–16) and HIGHWAVE (2019–26). F. Dias expands the project's scope by providing specific coastal and hydrodynamic applications that complement BRGM's expertise [25, 12]. His involvement establishes a direct connection to the geoscience community.

Consortium			
AgroParisTech			
Gloria Buritica	MCF	Task B (leader) and Task A (member)	20 p.m
Members tasks	PR	Tabea Rebafka (Task B member)	3 p.m
PhD student 1		Member tasks A, B	36 p.m
M2 intern 1		Member task A (real cases 1 and 3)	6 p.m
M2 intern 3		Member task A (real case 2)	6 p.m
BRGM			
Jérémy Rohmer	Research Eng.	Task A (leader)	4.60 p.m
Members tasks	Research Eng.	Anne Lemoine (real case 3), Rodrigo Pedreros (real cases 1 and 3) and Marc Peruzzetto (real case 2)	3.80 p.m
Centre Borelli			
Anne Sabourin	PR	Task C (leader) and Task A, B (member)	10 p.m
Members tasks	PR	Frédéric Dias (Task C member)	3 p.m
PhD student 2		Member Task C (real case 3)	36 p.m
M2 intern 2		Member Tasks A, C (real case 3)	6 p.m
		Total:	**131.9**

3.3 Management framework

One PhD student will be hired by AgroParisTech, and research visits at BRGM to meet with team members will allow progress in **Tasks A**. One objective of the thesis is to develop software containing the ML methodology for extremal regression, and to validate it on both simulation studies and real data applications. The student will produce an operational R and Python package to conduct the extreme impact studies of **Task A**. The second objective regarding the theoretical guarantees of the methods **Task B**, will progress with regular meetings with A. Sabourin in Centre Borelli. We highlight the geographical proximity of the two institutions, which will be beneficial for the project progress.

Hosted by Centre Borelli and supervised by A. Sabourin, the second PhD student will focus on the statistical learning and mathematical aspects of the project (**Tasks C1, C2**). While deeply rooted in theoretical foundations, this work will be driven by the practical applications developed in **Task A** and enriched through regular collaborations with applied profile team members J. Rohmer and F. Dias.

All along the project BRGM will play a central role in the discussions and contribute with field expertise and operational expertise from team members from the consortium. BRGM will more specifically support the project with: 1) the performance analysis of the new ML developments with confrontation to existing studies using 'classical' ML models; 2) feasibility analysis and transfer to operational studies (early warning systems and probabilistic hazard mapping). Several week-long visits to BRGM, located close to Paris in Orleans, are planned for the PhD student to guarantee the transfer knowledge regarding the data and the geophysical and weather phenomenon that it describes.

3.4 Institutional strategy

In 2023, AgroParisTech expressed the determination to foster risk adaptation research in his research strategy road map. The hosting laboratory MIA-Paris Saclay enhances connections between University Paris-Saclay and AgroParisTech-INRAE: two leading institutions in interdisciplinary research with an expertise in statistics/ML for addressing problems in life and environmental sciences. In addition, the cooperation between AgroParisTech and BRGM points to a key partnership between national institutions highly committed to natural risk prevention and Centre Borelli a laboratory in statistics.

Moreover, the problem tackled within this project is strongly related to two main operational activities at BRGM: 1. the development of prediction systems of natural hasards illustrated with: (i) Sirenes dedicated to the implementation of a forecasting and alert system for coastal hazards in Hauts-de-France; (ii) Tsunalogue dedicated to the developement of rapid prediction systems for Tsunami in Martinique in collaboration with the French Ministry for crisis management DGSCGC [42]; 2. The production of probabilistic hazard maps. See also an example for a municipality in Jura region.

In both situations, BRGM is currently working in research projects to explore the feasibility of using ML models like random forest, Gaussian process regression and neural networks as part of the predictive geosciences initiative. These modelling techniques don't tackle the problem of extremes. Therefore, the project is seen as a good opportunity to address in the short term the difficult question of predicting low likelihood, potentially high impact events. The project is a research exploratory project and depending on the outcomes, it can be the basis for future operational projects in the medium term aiming at modifying the current modelling and prediction practices. The long term strategy is to improve the prediction systems which is a key component towards the implementation of digital twins.

Since January 2026, Centre Borelli has hosted A. Sabourin as an associate researcher, with the strategic objective of establishing a dedicated research group, and ultimately a chair, focused on risks associated with extreme events in geosciences and climate science. This initiative is fully aligned with the laboratory's scientific policy and is actively supported by the ENS Paris-Saclay corporate foundation (*cellule mécénat*). The development of this chair is already underway, with Crédit Agricole Île-de-France currently funding a postdoctoral position (€75k), demonstrating strong institutional momentum and industrial commitment.

4 Expected outcomes of the project

From an applied perspective, the expected outcomes relate to incorporating ML methods for prediction of natural risks when dealing with events that are unlikely but have a significant impact. We also aim at improving trust in the use of these techniques by assessing the validity of the approaches used as well as the identification of the main drivers of the prediction errors of extremes.

Open-source publications in top ML and statistics journals and conferences, with the theoretical guarantees of the methods proposed, together with communication of the findings through national and international workshops and conferences is expected. We also aim to publish in open-source weather and geophysical journals with the details of the implementation of the methodology on case studies oriented towards the community of future users to demonstrate the usefulness of the novel tools. Dissemination will mainly include: (1) national and international journal papers and conferences, (2) data, code and packages via a Github/lab repository, (3) a project website hosted by BRGM based on open-source tools (4) initial and closing workshops (5) communication through dedicated networks of PEPR IRIMA (in particular IRIMONT, IRICOT, Outre-Mer) and Maths-Vives (in particular CLIMATHS, COMPLEXFLOWS, and HYDRAUMATH) and PEPR TRACCS (EXTENDING) (6) communication toward the general public.

Finally, an opening workshop will allow to involve actors working on natural risks and statistics for risk quantification and share/identify the difficulties and main challenges in the intersection of both fields. The project will also close with a final workshop to ensure knowledge transferability.

5 Funding justification

Budget: 507.4 kEur requested funding		
AgroParisTech	Justification	**198K**
1×PhD grant	36 months	125
2×M2 intern grants	6 months each	7.5
Instrument and material costs	A computer for the PhD student	2.5
Travel / conferences	2 kEur / person / year	20
Launching workshop	Transfer of case study and current operation status	10
Overheads costs		33
BRGM	Justification	**122.2K**
Permanent staff	4 persons for support / recommendations on the 3 real cases, feasability assessment for operational studies	79
Travel / conferences	0.5 kEur / person / year	10
Final workshop	Transfer to risk managers and early-warning operators (i.e., presenters' invitation, meals, venue hire)	13
Overheads costs		20
Borelli Center	Justification	**187.2K**
1×PhD grant	36 months	129
1× M2 intern grants	6 months	4.5
Instrument and material costs	A computer for the PhD student	2.5
Travel / conferences	2 kEur / person / year	20
Overheads costs		31.3
	Total:	**507.4**

The funding we request allows us to hire two PhD students, and three master interns as well as the necessary equipment for both PhD students to conduct their research. We also include the budget mobility to finance the visits among all collaborators from Table 1. These visits have the purpose of enhancing the exchanges between collaborators and they guarantee a follow-up of the PhD students as part of the responsibilities described above. Finally, the expertise services of BRGM including their supervising responsibilities will be covered with the budget allocated to the external expenses. The budget also covers the opening and the closing workshops to foster the transfer knowledge of the project outcomes.

6 Cited references

References

[1] Mickael Accensi, Matias Alday, Christophe Maisondieu, Nicolas Raillard, David Darbynian, Chris Old, Brian Sellar, Olivia Thilleul, Yves Perignon, Gregory Payne, et al. Resourcecode framework: A high-resolution wave parameter dataset for the european shelf and analysis toolbox. In 14th European Wave and Tidal Energy Conference, EWTEC 2021, pages 2182–1. EWTEC, 2021.

[2] Vineeth N Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. Conformal prediction for reliable machine learning: theory, adaptations and applications. Elsevier, 2014.

[3] Yann Balouin, Ywenn de la Torre, Rodrigo Pedreros, Andrea Gilberto Filippini, Thibault Laigre, Noémie Videau, Christophe Blazy, Eric Abadie, Thibault Catry, Julie Deshayes, et al. Carib-coast project: Caribbean network for prevention of coastal risks related to climate change. The Proceedings of the Coastal Sediments 2023 In 5 Volumes, 2023.

[4] Bojan Basrak and Johan Segers. Regularly varying multivariate time series. Stochastic Processes and their Applica, 2011.

[5] Thomas B Berrett, Ioannis Kontoyiannis, and Richard J Samworth. Optimal rates for independence testing via u-statistic permutation tests. The Annals of Statistics, 49(5):2457–2490, 2021.

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

[7] **Buriticá, Gloria**, Nicolas Meyer, Thomas Mikosch, and Olivier Wintenberger. Some variations on the extremal index. Journal of mathematical sciences, 273(5):687–704, 2023.

[8] **Buriticá, Gloria** and Philippe Naveau. Stable sums to infer high return levels of multivariate rainfall time series. Environmetrics, 34(4):e2782, 2023.

[9] **Buriticá, Gloria** and Olivier Wintenberger. On the asymptotics of extremal ℓ_p-blocks cluster inference. Extremes, 2025.

[10] Victor Chernozhukov and Alexandre Belloni. L1-penalized quantile regression in high dimensional sparse models. 2009.

[11] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):6614–6630, 2021.

[12] Colm Clancy, John O'Sullivan, Conor Sweeney, **Dias, Frédéric**, and Andrew C Parnell. Spatial bayesian hierarchical modelling of extreme sea states. Ocean Modelling, 107:1–13, 2016.

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

[14] Stéphan Clémençon, Hamid Jalalzai, Stéphane Lhaut, **Sabourin, Anne**, and Johan Segers. Concentration bounds for the empirical angular measure with statistical learning applications. Bernoulli, 29(4):2797–2827, 2023.

[15] Stephan Clémençon, Myrto Limnios, and Nicolas Vayatis. A bipartite ranking approach to the two-sample problem. Electronic Journal of Statistics, 19(1):2733–2779, 2025.

[16] Stephan Clémençon and **Sabourin, Anne**. Weak signals and heavy tails: learning theory meets extreme value analysis. Extremes, pages 1–43, 2026.

[17] Daniel Cooley, Richard A Davis, and Philippe Naveau. Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. 2012.

[18] Miguel De Carvalho, Alina Kumukova, and Gonçalo Dos Reis. Regression-type analysis for multivariate extreme values. Extremes, 25(4):595, 2022.

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

[20] Clément Dombry and Ahmed Zaoui. Distributional regression: Crps-error bounds for model fitting, model selection and convex aggregation. Advances in Neural Information Processing Systems, 37:32678–32704, 2024.

[21] Holger Drees and **Sabourin, Anne**. Principal component analysis for multivariate extremes. Electronic Journal of Statistics, 15:908–943, 2021.

[22] John H. J. Einmahl, Fan Yang, and Chen Zhou. Testing the multivariate regular variation model. Journal of Business & Economic Statistics, 39:907 – 919, 2018.

[23] Sebastian Engelke, Nicola Gnecco, and **Sabourin, Anne**. Extrapolation in statistical learning with extreme value theory. arXiv preprint arXiv:2605.01909, 2026.

[24] Sebastian Engelke and Jevgenijs Ivanovs. Sparse structures for multivariate extremes. Annual Review of Statistics and Its Application, 8(1):241–270, 2021.

[25] Francesco Fedele, Joseph Brennan, Sonia Ponce de León, John Dudley, and **Dias, Frédéric**. Real world ocean rogue waves explained without the modulational instability. Scientific reports, 6(1):27715, 2016.

[26] Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. Bernoulli, 29(1):1–23, 2023.

[27] United Nations Office for Disaster Risk Reduction. Global Assessment Report on Disaster Risk Reduction 2025: Resilience Pays: Financing and Investing for Our Future. Stylus Publishing, LLC, 2025.

[28] Laurent Gardes. Tail dimension reduction for extreme quantile estimation. Extremes, 21(1):57–95, 2018.

[29] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87(4):1100–1126, 2025.

[30] Stéphane Girard and Cambyse Pakzad. Functional extreme-pls. arXiv preprint arXiv:2410.05517, 2024.

[31] Christophe Giraud. Introduction to high-dimensional statistics. Chapman and Hall/CRC, 2021.

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

[33] Yuri Goegebeur, Armelle Guillou, and Jing Qin. Marginal expected shortfall risk measure for time series: Y. goegebeur et al. Statistical Inference for Stochastic Processes, 28(3):16, 2025.

[34] Nicolas Goix, **Sabourin, Anne**, and Stéphan Clémençon. Sparse representation of multivariate extremes with applications to anomaly ranking. In Artificial intelligence and statistics, pages 75–83. PMLR, 2016.

[35] Nicolas Goix, **Sabourin, Anne**, and Stephan Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. Journal of Multivariate Analysis, 161:12–31, 2017.

[36] Seth Guikema. Artificial intelligence for natural hazards risk analysis: Potential, challenges, and research needs. Risk Analysis, 40(6):1117–1123, 2020.

[37] Janet E Heffernan and Jonathan A Tawn. A conditional approach for multivariate extreme values (with discussion). Journal of the Royal Statistical Society Series B: Statistical Methodology, 66(3):497–546, 2004.

[38] Christopher David Jack, Richard Jones, Laura Burgin, and Joseph Daron. Climate risk narratives: An iterative reflective process for co-producing and integrating climate knowledge. Climate Risk Management, 29:100239, 2020.

[39] Hamid Jalalzai, Stephan Clémençon, and **Sabourin, Anne**. On binary classification in extreme regions. Advances in Neural Information Processing Systems, 31, 2018.

[40] Roger Koenker. Quantile regression. Econometric Society Monographs, Cambridge University Press, Cambridge, 2005.

[41] Rafal Kulik and Philippe Soulier. Heavy-Tailed Time Series. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 1 edition, 2020.

[42] Anne Lemoine, Rodrigo Pedreros, Yoann Legendre, Romain Guidez, Caterina Negulescu, Sylvestre Le Roy, **Rohmer, Jérémy**, Aude Nachbaur, Frédéric Dondin, Cécile Gracianne, et al. Tsunalogue: Elaboration d'un démonstrateur d'analyse simple pour le déclenchement rapide d'alerte tsunami en martinique. Technical report, 2025.

[43] Baptiste Leroux, Clément Dombry, and **Sabourin, Anne**. Out-of-distribution generalization of quantile regression with heavy tailed inputs: an svm approach. arXiv preprint arXiv:2606.00265, 2026.

[44] Stéphane Lhaut, **Sabourin, Anne**, and Johan Segers. Uniform concentration bounds for frequencies of rare events. Statistics & Probability Letters, 189:109610, 2022.

[45] Myrto Limnios and Stéphan Clémençon. On ranking-based tests of independence. In International Conference on Artificial Intelligence and Statistics, pages 577–585. PMLR, 2024.

[46] Shan Lin, Zenglong Liang, Hongwei Guo, Quanke Hu, Xitailang Cao, and Hong Zheng. Application of machine learning in early warning system of geotechnical disaster: a systematic and comprehensive review. Artificial Intelligence Review, 58(6):168, 2025.

[47] Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. The Annals of Statistics, 51(2):738–761, 2023.

[48] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Yang Chen, Leah Goldfarb, Melissa I Gomis, et al. Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, 2(1):2391, 2021.

[49] Marco Avella Medina, Richard A Davis, and Gennady Samorodnitsky. Spectral learning of multivariate extremes. Journal of Machine Learning Research, 25(124):1–36, 2024.

[50] Nicolas Meyer and Olivier Wintenberger. Sparse regular variation. Advances in Applied Probability, 53(4):1115–1148, 2021.

[51] Thomas Mikosch and Olivier Wintenberger. Extreme value theory for time series. Models With, 2024.

[52] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. Advances in neural information processing systems, 21, 2008.

[53] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. Journal of Machine Learning Research, 11(2), 2010.

[54] 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.

[55] Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In International Conference on Machine Learning, pages 17517–17530. PMLR, 2022.

[56] Marc Peruzzetto, Clara Levy, Bastien Colas, Yannick Thiery, Séverine Bernardie, Benoit Vittecoq, Aude Nachbaur, Geoffrey Aertgeerts, Anne Mangeney, and Stéphane Buschaert. Simulations numériques pour la cartographie de l'aléa gravitaire. In Assises Nationales des Risques Natuels-Programme Scientifique, 2025.

[57] Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. Advances in Statistical Climatology, Meteorology and Oceanography, 11(1):23–58, 2025.

[58] Sidney I Resnick. Heavy-tail phenomena: probabilistic and statistical modeling. Springer, 2007.

[59] **Rohmer, Jérémy**, Andrea G Filippini, and Rodrigo Pedreros. Combining 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.

[60] **Rohmer, Jérémy**, Marie Rousseau, Anne Lemoine, Rodrigo Pedreros, J Lambert, and A Benki. Source characterisation by mixing long-running tsunami wave numerical simulations and historical observations within a metamodel-aided abc setting. Stochastic environmental research and risk assessment, 32(4):967–984, 2018.

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

[62] Robert Šakić Trogrlić, Marc Van Den Homberg, Mirianna Budimir, Colin McQuistan, Alison Sneddon, and Brian Golding. Early warning systems and their role in disaster risk reduction. In Towards the "perfect" weather warning: bridging disciplinary gaps through partnership and communication, pages 11–46. Springer, 2022.

[63] Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87(3):653–677, 2025.

[64] Roland B Sogan, **Rebafka, Tabea**, and Fanny Villers. A bayesian framework with adaptive elastic nets for the inference of gaussian graphical models. arXiv preprint arXiv:2604.18042, 2026.

[65] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288, 1996.

[66] Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. Journal of Machine Learning Research, 25(216):1–109, 2024.

[67] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic Learning in a Random World. Springer, New York, 2005.

[68] Vladimir Vovk, Alexander Gammerman, and Craig Saunders. Machine learning applications of algorithmic randomness. In Proceedings of the Sixteenth International Conference on Machine Learning, pages 444–453, 1999.

[69] Jennifer L Wadsworth and Ryan Campbell. Statistical inference for multivariate extremes via a geometric approach. Journal of the Royal Statistical Society Series B: Statistical Methodology, 86(5):1243–1265, 2024.

[70] Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. Annual Review of Statistics and Its Application, 13, 2025.

[71] Phyllis Wan and RA Davis. Threshold selection for multivariate heavy-tailed data. Extremes, 22, 2018.

[72] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(4):4396–4415, 2022.

7. Project leader CV

Personal details					
Gender *(Woman, man, prefers not to answer)*	Woman				
Name and first name	Gloria Buriticá				
Nationality / Residence	Colombian / France				
Current position / Poste actuel					
Function	Associate professor (MCF)				
French public organisation(s) / Organisme(s) public(s) français					
Code RNSR	Organisation	Laboratory	Unit code	Postcode	Town
201119642G	AgroParisTech	MIA Paris- Saclay	UMR 0518	91123	Palaiseau
Other activities (*Services, commitment in scientific events, scientific editing, scientific reviewing activities)*					
- Workshop organization: 'BIRS 26w565-Catastrophic Events in the Complex Word: Mathematics and Statistics of extremes in the Age of Machine learning' in BIRS, Banff, Canada (to come in Aug. 2026). 'Causality in Extremes' at the University of Geneva 2024. - Referee in: Annals of Statistics, Bernoulli, Extremes, Journal of Applied Probability, Journal of Machine Learning Research, Test journal.					
Pedagogical and administrative responsability					
- **Masters' internship supervision** : (4 successful master reports, 2 ongoing master internships) - Hakim Barlier, 2026 (M2 ongoing), AgroParisTech. Francis Jégou, 2026 (M2 ongoing), AgroParisTech. Philip Boustani, 2024, M2 U. Geneva (currently pursuing a PhD at LMU), Luciana D'souza, 2024, M2 U. Geneva. Marc Gonzalez, 2023, M2 U. Geneva. Alexandre Mansire, 2023, Sorbonne Université (M1). - Member of the selection committee for incoming mobility candidates at AgroParisTech					

Previous positions				
Function / Titre				
Start date	End date	Town	Organisation	Function
Oct. 2022	Sep. 2024	Geneva	University of Geneva	Postdoctoral researcher

Education / Formation supérieure
2019-2022 PhD Statistics, Sorbonne Université, Laboratoire de Probabilités, Statistiques et Modélisation LPSM
2017-2019 Master in probability and statistics, Sorbonne Univesité, LPSM
2013-2017 Undergraduate in Mathematics, Universidad de los Andes, Bogotá

Scientific productions / Productions scientifiques

	5 publications / 5 publications	What is the major contribution of this publication? / Quel est l'apport majeur de cette publication ?
1	Buriticá, G., & Engelke, S. (2024). Progression: an extrapolation principle for regression. https://doi.org/10.48550/arXiv.2410.23246	-Design and implement *Progression* algorithm for extremal regression with a univariate covariate. -Derive error bounds for the regression extrapolation function coded in the *Progression* algorithm.
2	Buriticá, G., & Naveau, P. (2022). Stable sums to infer high return levels of multivariate rainfall time series, *Environmetrics.* e2782. https://doi.org/10.1002/env.2782	-Propose and implement a methodology for high return levels (quantiles) estimation for dependent time series relying on α-stable distribution. -Case study: French precipitation weather stations' network
3	Buriticá, G., Hentschel, M., Pasche, O. *et al.* Modeling extreme events: Univariate and multivariate data-driven approaches. *Extremes* (2024). https://doi.org/10.1007/s10687-024-00499-9	-Present a method for high quantile inference in the setting of asymmetric loss risk evaluation. -Participation and 1st place (C2) in the Extreme Value Analysis EVA conference data challenge 2023
4	G.Buriticá, T. Mikosch, O. Wintenberger (2023). Large deviations of ℓp −blocks of regularly varying time series and applications to cluster inference. Stochastic Processes and their Applications. 161, 68–101. https://doi.org/10.1016/j.spa.2023.03.013	-Propose a family of estimators for extremal cluster inference, and prove its asymptotic consistency. -The method is based on the extremal properties of consecutive observations studied in the article via large deviations' results for sums of heavy-tailed time series.
5	Buriticá, G., Meyer, N., Mikosch, T. *et al.* Some Variations on the Extremal Index. *J Math Sci* **273**, 687–704 (2023). https://doi.org/10.1007/s10958-023-06533-8	-Modeling extremal clustering behavior of heavy-tailed time series using point processes. -Propose a method for inference of the extremal index: a key quantity for high quantile estimation.
	Data set, software, source code, data paper,	Brief description
1	R package *Progression* sur gitHub https://github.com/GBuritica/progression_r	Tools for out-of-domain regression based on extreme value theory coupled with adaption of ML algorithms
1	R package tsExtremes sur gitHub https://gburitica.github.io/tsExtremes/	Tools with tutorial for analysing the extreme behavior of heavy-tailed time series: shape parameter estimation, extremal index inference and cluster lengths estimates.

8. Linked projects / Projets liés

BRGM will rely on different past and on-going research and applied projects to define the application test cases described in **Task A**, namely:

• Real case 1 based on the Interreg-CaribCoast projectCarib-Coast project finalised in 2022 [3], and further used in PEPR IRIMA Outre Mer.

• Real case 2 based on the recently launched ANR JCJC PROBTOOL project as well as PEPR IRIMAIRIMONT.

• Real case 3 based on the AMI ANR TANDEM project, and may be extended via the on-going project "tsunalogue" [43].

Methodologically, this proposal builds upon the broader framework of the ANR project EXSTA (2023–2027, P.I.: A. Sabourin). While EXSTA establishes key foundations for penalized empirical risk minimization under extreme covariates—notably via strict i.i.d. guarantees as in [43]—it leaves critical theoretical and operational bottlenecks open. This project introduces a major departure from EXSTA's core theoretical scope by shifting from i.i.d. settings to non-i.i.d. temporal dependencies, dimension reduction of predictors, and distribution-free uncertainty quantification (Tasks B and C). Bridging advanced mathematical statistics with multidisciplinary challenges, we translate these new probabilistic guarantees under dependence into a rigorous, deployable toolkit for supervised extremal regression in applied fields