

# Introductory statistical ideas for IOGP: model fitting

Philip Jonathan and Gemini AI

September 2026

## 1 Maximum likelihood estimation of generalised Pareto parameters

For a random variable  $X$  exceeding a high threshold  $u$ , let the exceedance data be given by the sample  $\{x_i\}_{i=1}^n$ , where each  $y_i = x_i - u > 0$ . The Generalised Pareto Distribution (GPD) has a probability density function for exceedances given by:

$$f(y_i | \sigma, \xi) = \frac{1}{\sigma} \left(1 + \xi \frac{y_i}{\sigma}\right)^{-1/\xi-1} \quad (1)$$

where the parameter domains are: scale parameter  $\sigma > 0$  and shape parameter  $\xi \in \mathbb{R}$  (with the support constraint  $1 + \xi \frac{y_i}{\sigma} > 0$  for all  $i$ ).

To estimate the parameters  $\hat{\xi}$  and  $\hat{\sigma}$  via Maximum Likelihood Estimation (MLE), we formulate the total sample likelihood  $\mathcal{L}(\sigma, \xi)$  as the joint probability density of the independent and identically distributed exceedances:

$$\mathcal{L}(\sigma, \xi) = \prod_{i=1}^n f(y_i | \sigma, \xi) = \prod_{i=1}^n \frac{1}{\sigma} \left(1 + \xi \frac{y_i}{\sigma}\right)^{-1/\xi-1} \quad (2)$$

The maximum likelihood estimates  $(\hat{\sigma}, \hat{\xi})$  are obtained by maximizing  $\mathcal{L}(\sigma, \xi)$  over the parameter space  $\sigma > 0$  and  $1 + \xi \frac{y_{(n)}}{\sigma} > 0$  (where  $y_{(n)}$  is the maximum exceedance).

## 2 Non-parametric bootstrap uncertainty quantification

To quantify the uncertainty associated with the parameter estimates  $\hat{\xi}$  and  $\hat{\sigma}$ , we apply a non-parametric bootstrap resampling scheme.

### Algorithm for non-parametric bootstrap uncertainty quantification:

1. Given the original sample of threshold exceedances  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$  of size  $n$ .
2. For bootstrap iteration  $b = 1, 2, \dots, B$  (where  $B$  is large, e.g.,  $B = 1000$ ):
  - (a) Draw a resample  $\mathbf{y}^{(b)} = \{y_1^*, y_2^*, \dots, y_n^*\}$  of size  $n$  uniformly at random with replacement from the original sample  $\mathbf{y}$ .
  - (b) Maximize the sample likelihood  $\mathcal{L}(\sigma, \xi)$  on the resample  $\mathbf{y}^{(b)}$  to obtain bootstrap parameter estimates  $(\hat{\sigma}^{(b)}, \hat{\xi}^{(b)})$ .
3. Compute the empirical standard errors and percentile confidence intervals from the empirical distributions of  $\{\hat{\sigma}^{(b)}\}_{b=1}^B$  and  $\{\hat{\xi}^{(b)}\}_{b=1}^B$ .

### 3 Cross-validation for predictive performance

To evaluate the predictive performance of a fitted GPD model without overfitting, we implement a K-fold cross-validation scheme.

**Algorithm for cross-validation of predictive performance:**

1. Partition the dataset of threshold exceedances into  $K$  roughly equal-sized subsets (folds)  $S_1, S_2, \dots, S_K$ .
2. For each fold  $k = 1, 2, \dots, K$ :
  - (a) Train the GPD model by maximizing the sample likelihood  $\mathcal{L}(\sigma, \xi)$  on all data except fold  $k$  (i.e., training set  $\bigcup_{j \neq k} S_j$ ), yielding parameter estimates  $\hat{\sigma}_{(-k)}, \hat{\xi}_{(-k)}$ .
  - (b) Evaluate the predictive performance on the held-out fold  $S_k$  using a proper scoring rule, such as the negative log-likelihood score:

$$\text{Score}_k = - \sum_{y_i \in S_k} \ln f(y_i \mid \hat{\sigma}_{(-k)}, \hat{\xi}_{(-k)}) \quad (3)$$

3. Aggregate the cross-validation score across all folds:  $\text{CVScore} = \sum_{k=1}^K \text{Score}_k$ , representing out-of-sample predictive adequacy.

### 4 Cross-validation for threshold selection

Selecting the optimal threshold  $u$  balances bias (choosing too low, violating asymptotic GPD assumptions) and variance (choosing too high, leaving too few exceedances).

**Algorithm for cross-validation threshold selection:**

1. Define a grid of candidate thresholds  $u_1, u_2, \dots, u_m$  corresponding to high sample quantiles (e.g., 80<sup>th</sup> to 95<sup>th</sup> percentiles).
2. For each candidate threshold  $u_j$ :
  - (a) Extract the set of exceedances  $Y(u_j) = \{x_i - u_j : x_i > u_j\}$ .
  - (b) Perform K-fold cross-validation on  $Y(u_j)$  as outlined in Section 3 to compute the out-of-sample negative log-likelihood score  $\text{CVScore}(u_j)$ .
3. Select the optimal threshold  $u^*$  that minimizes  $\text{CVScore}(u)$ :

$$u^* = \arg \min_{u_j} \text{CVScore}(u_j) \quad (4)$$

### 5 Model ensembles and model averaging

Relying on a single model structure, bootstrap sample, or threshold can lead to structural uncertainty. Combining an ensemble of models provides more robust and stable predictions.

**Motivation for model averaging:**

- **Threshold uncertainty:** Different thresholds yield slightly different tail estimates; averaging across models fitted at multiple thresholds accounts for threshold selection risk.
- **Bootstrap uncertainty:** Averaging predictions across bootstrap resamples stabilizes variance and reduces bias in finite samples.

- **Structure uncertainty:** Combining different covariate relationships or parametric families captures structural model uncertainty.

**Algorithm for model averaging prediction:**

1. Fit a set of  $M$  distinct GPD models (varying across thresholds, bootstraps, or covariate formulations), yielding parameter sets  $\{(\hat{\sigma}_m, \hat{\xi}_m)\}_{m=1}^M$ .
2. Assign model weights  $w_m$  based on information criteria (e.g., AIC/BIC weights) or out-of-sample cross-validation likelihood scores, such that  $\sum_{m=1}^M w_m = 1$  and  $w_m \geq 0$ .
3. Compute the model-averaged return level or tail probability  $\hat{Q}(p)$  as the weighted sum of individual model predictions:

$$\hat{Q}(p) = \sum_{m=1}^M w_m \hat{Q}_m(p) \quad (5)$$

## 6 Penalised likelihood for non-stationary spline models

When tail parameters vary smoothly with a covariate  $\Theta$  (e.g., storm direction or season), we model non-stationarity using regression splines.

Let the linear predictors for the transformed parameters be represented using B-spline bases  $B$ :

$$\boldsymbol{\eta}_\xi = \mathbf{B}\boldsymbol{\beta}_\xi, \quad \boldsymbol{\eta}_\sigma = \mathbf{B}\boldsymbol{\beta}_\sigma \quad (6)$$

where  $\mathbf{B}$  is the design matrix of B-spline basis functions evaluated at covariate values, and  $\boldsymbol{\beta}_\xi, \boldsymbol{\beta}_\sigma$  are spline coefficient vectors. Domains: spline coefficients  $\boldsymbol{\beta} \in \mathbb{R}^d$ , with smoothing penalties  $\lambda_\xi \geq 0, \lambda_\sigma \geq 0$ .

To prevent overfitting and ensure smooth variations, we maximize the penalized sample likelihood  $\mathcal{L}_p$ :

$$\mathcal{L}_p(\boldsymbol{\beta}_\xi, \boldsymbol{\beta}_\sigma) = \left( \prod_{i=1}^n f(y_i | \sigma(\theta_i), \xi(\theta_i)) \right) \times \exp \left( -\frac{\lambda_\xi}{2} \boldsymbol{\beta}_\xi^T \mathbf{S} \boldsymbol{\beta}_\xi - \frac{\lambda_\sigma}{2} \boldsymbol{\beta}_\sigma^T \mathbf{S} \boldsymbol{\beta}_\sigma \right) \quad (7)$$

where  $\mathbf{S}$  is a roughness penalty (based on squared differences for P-splines).

**Algorithm for fitting a non-stationary spline GPD model:**

1. Collect paired exceedances and covariate observations  $\{(y_i, \theta_i)\}_{i=1}^n$ .
2. Construct B-spline basis matrices  $\mathbf{B}$  over the domain of covariate  $\Theta$ .
3. Initialize smoothing parameters  $\lambda_\xi$  and  $\lambda_\sigma$  using cross-validation.
4. Maximize the penalized sample likelihood  $\mathcal{L}_p$  iteratively (using penalized iteratively reweighted least squares, backfitting or Newton-Raphson optimization) to estimate  $\hat{\boldsymbol{\beta}}_\xi$  and  $\hat{\boldsymbol{\beta}}_\sigma$ .
5. Evaluate covariate-varying tail parameters  $\hat{\xi}(\theta) = \mathbf{B}(\theta)\hat{\boldsymbol{\beta}}_\xi$  and  $\hat{\sigma}(\theta) = \exp(\mathbf{B}(\theta)\hat{\boldsymbol{\beta}}_\sigma)$ .

## 7 Bayesian equivalence of spline smoothness constraints

From a Bayesian perspective, introducing a penalty term in penalized likelihood optimization is mathematically equivalent to placing a prior distribution on the spline coefficients.

Consider the penalized sample likelihood from Section 6. Taking the logarithm of the penalty term reveals its correspondence to a Gaussian prior distribution on the spline coefficients  $\boldsymbol{\beta}_\xi \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{S}^{-1})$ , where the smoothing parameter corresponds to the precision ratio  $\lambda_\xi = 1/\tau^2$ .

Thus, maximizing the penalized sample likelihood is equivalent to finding the Maximum A Posteriori (MAP) estimate under a Bayesian hierarchical model:

$$\mathcal{L}_p(\boldsymbol{\beta}) \propto \mathcal{L}(\boldsymbol{\beta}) \times p(\boldsymbol{\beta} \mid \lambda) \quad (8)$$

where  $\mathcal{L}(\boldsymbol{\beta})$  is the standard sample likelihood  $\prod_{i=1}^n f(y_i \mid \boldsymbol{\beta})$ , and  $p(\boldsymbol{\beta} \mid \lambda) \propto \exp(-\frac{\lambda}{2} \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta})$  acts as the prior probability density function for the parameters.

## 8 Prompt Script Reproduction

The following text block reproduces the exact prompt script used to generate this document, formatted in verbatim style with word-wrapping enabled.

```
You are a university teacher of applied statistics to engineering students.

Your students have a rudimentary understanding of probability distributions and
statistical modelling.

The students need a basic understanding of statistical distributions, methods for
describing the TAILS of those distributions, and methods for fitting models to
data.

You are going to produce a LaTeX report in multiple sections.

In producing the report:
- Please confirm that all LaTeX code generated is valid, and that there are no minor
  typographical errors.
- Please provide full step by step mathematical derivations when asked to "demonstrate
  analytically" are requested.
- Please provide full step by step algorithms when asked to "Give algorithm".
- With every concept introduced, please provide a simple equation explaining the
  concept, NOT JUST WORDS.
- Please define the domains of all parameters in statistical models
- Please use sentence case for all titles
- Please use the report title "Introductory statistical ideas for IOGP masterclass"
  and sub-title "Stats for Met-ocean"
- Please state that the authors are "Philip Jonathan and Gemini AI"
- Please make sure that LaTeX is compatible with pdftex

Please provide a LaTeX report with the following sections:

1. An introduction to maximum likelihood estimation of the generalised Pareto tail
   parameters  $\hat{\xi}$  and  $\hat{\sigma}$  from a sample  $\{x_i\}_{i=1,2,\dots,n}$  of random
   variable  $X$  above threshold  $u$ . Please use  $\mathcal{L}$  to describe SAMPLE LIKELIHOOD
   and  $\ell$  for sample log-likelihood
- please describe in term of the SAMPLE LIKELIHOOD not the sample log-likelihood
- Give equation for SAMPLE LIKELIHOOD

2. Give algorithm for a simple non-parametric bootstrap scheme to provide
   quantification of uncertainty for  $\hat{\xi}$  and  $\hat{\sigma}$ 

3. Give algorithm for a cross-validation scheme to evaluate the PREDICTIVE performance
   of the fitted generalised Pareto model

4. Give algorithm for a cross-validation scheme to evaluate tuning parameters (such as
   threshold) for optimal PREDICTIVE performance of the fitted generalised Pareto
   model
```

5. Provide motivation for why an ensemble of models can be used to provide predictions in place of a single model
  - provide examples of different reasons to model average (different thresholds, different bootstraps, different model structures)
  - Give algorithm
6. Provide motivation for a penalised likelihood optimisation to estimate spline-based representations of variation of  $\xi$  and  $\sigma$  with a covariate  $\Theta$  (such as storm direction, in the case that  $X$  corresponds to storm peak significant wave height)
  - Give equation for spline representation of parameter  $\eta_{\xi} = B \beta_{\xi}$  and  $\eta_{\sigma} = B \beta_{\sigma}$  for common spline bases  $B$  evaluated on some index set on the covariate domain.
  - please describe in term of the SAMPLE LIKELIHOOD not the sample log-likelihood
  - Give algorithm for fitting of non-stationary spline generalised Pareto model
7. Give equation illustrating the equivalence of the specified spline smoothness constraint in penalised likelihood optimisation, and a prior for spline parameters from a Bayesian analysis perspective
  - please describe in term of the SAMPLE LIKELIHOOD not the sample log-likelihood
8. Reproduce this prompt script explicitly, so that readers can see how Gemini AI was used to generate it.
  - Please use word wrapping to avoid overly long lines not formatting properly in the pdf.
  - please use VERBATIM to ensure that the prompt script is properly word-wrapped in the eventual pdf.
  - Please confirm that the .tex file compiles correctly