Diagnostic Tools for Non-stationary Extreme Value Models

E. Mackay J. Richards P. Jonathan

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

February 18, 2026

Motivation

- Non-stationary extreme value models arise in many applied domains, e.g., environmental sciences, public health, finance.
- Reliable extrapolation requires careful assessment of model fit
- Standard diagnostics pooled information across covariate space - hide problems!
- Goal: diagnostics that assess fit *locally* and *globally*

Objectives of Diagnostics

1. Assess overall goodness-of-fit
2. Identify regions of covariate space with poor fit
3. Facilitate comparison between competing models

Stationary Diagnostics Review

- Sample of observations $\{y_1, ..., y_n\}$ of random variable Y with distribution function F_Y
- Used to estimate a model with distribution function $\hat{F}_Y$
- (PIT) if $\hat{F}_Y = F_Y$, the values $\{\hat{F}_Y(y_1), ..., \hat{F}_Y(y_n)\} \sim Unif(0, 1)$.

Let $y_{(1)} \leq \cdots \leq y_{(n)}$ denote the ordered sample, and assign each $y_{(k)}$ an empirical non-exceedance probability p_k. The definition of p_k varies between practitioners, with $p_k = (k - 0.5)/n$ or $p_k = k/(n + 1)$ being common choices.

Stationary Diagnostics Review

A PP plot consists of the points

$$\left\{\left(p_k,\, \hat{F}_Y(y_{(k)})\right) : k = 1, ..., n\right\},$$

and a QQ plot consists of the points

$$\left\{\left(\hat{F}_Y^{-1}(p_k),\, y_{(k)}\right) : k = 1, ..., n\right\},$$

where $\hat{F}_Y^{-1}$ is the model's quantile function.
A transformed QQ plot consists of the points

$$\left\{\left(F_0^{-1}(p_k),\, F_0^{-1}\left(\hat{F}_Y(y_{(k)})\right)\right) : k = 1, ..., n\right\}.$$

Non-stationary Diagnostics Review

- Sample of observations $\{(y_i, \mathbf{x}_i)\}$ of random variable Y with conditional distribution function $F_{Y|\mathbf{X}}$ for $\mathbf{X} \in \mathcal{D} \subset \mathbb{R}^p$.
- Used to estimate a model with distribution function $\hat{F}_{Y|X}$
- (PIT) if $\hat{F}_{Y|\mathbf{X}} = F_{Y|\mathbf{X}}$, the independent values $\{\hat{F}_{Y|\mathbf{X}}(y_1|\mathbf{x}_1), ..., \hat{F}_{Y|\mathbf{X}}(y_n|\mathbf{x}_n)\} \sim Unif(0, 1)$.

Then, proceed as normal (using e.g., transformed QQ plot).

Challenges in Non-stationary Settings

Problems:

1. No local information - could create B plots by binning covariate space $\mathcal{D}$ into overlapping bins, $\mathcal{B}_1, \ldots, \mathcal{B}_B$, but difficult to visualise for large B + how to handle different n per bin?
2. One realisation of plot - cannot assess uncertainty or perform testing easily (without laborious bootstrap - not great if using AI!);

It would be desirable to have visual diagnostics which:

1. Summarise information about the fit of a model across multiple local regions of the covariate domain;
2. Are invariant to the shapes and scales of the distribution tails in each local region;
3. Are invariant to the sample size in each local region;
4. Provide a visual assessment of the significance of deviations between the model and observations.

Pooled Exponential QQ Plot

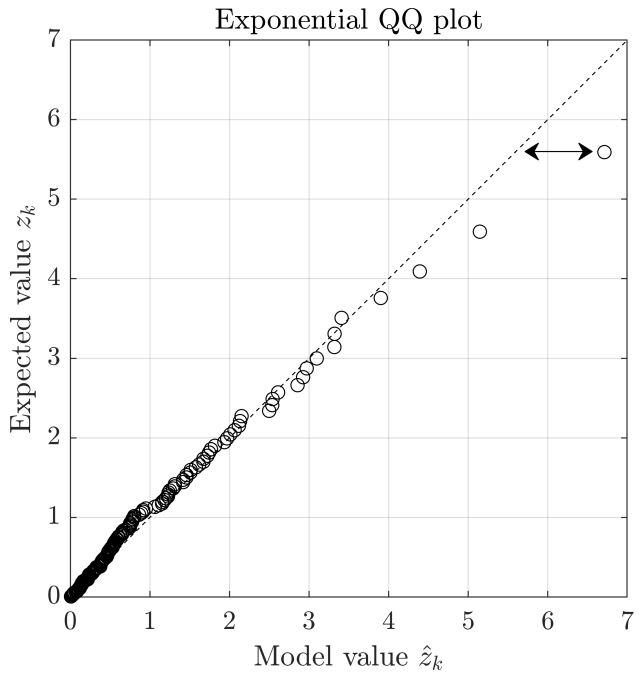
Suppose that we have a sequence of independent random pairs $\{(Y_k, \boldsymbol{X}_k)\}_{1:n}$, which could be over either a local region or global range of the covariate $\boldsymbol{X}$.

Define $P_k = F_{Y|\boldsymbol{X}}(Y_k|\boldsymbol{X}_k)$, $Q_k = 1 - P_k$, and $Z_k = -\log(Q_k)$.

Since $P_k \sim \mathrm{U}(0,1)$, we have $Z_k \sim \mathrm{Exp}(1)$. Denote the ordered exceedance probabilities $Q_{(1)} \leq \cdots \leq Q_{(n)}$ and exponential order statistics $Z_{(1)} \geq \cdots \geq Z_{(n)}$.

For a sample of observations $\{(y_i, \boldsymbol{x}_i)\}_{1:n}$ and a model $\hat{F}_{Y|\boldsymbol{X}}$, define $\hat{q}_k = 1 - \hat{F}_{Y|\boldsymbol{X}}(y_k|\boldsymbol{x}_k)$. Denote the ordered values $\hat{q}_{(1)} \leq \cdots \leq \hat{q}_{(n)}$, and define $\hat{z}_k = -\log(\hat{q}_{(k)})$. The pooled exponential QQ plot is a scatter plot of $\{(\hat{z}_k, z_k) : k = 1, ..., n\}$, where z_k is (in our case) $\mathbb{E}[Z_{(k)}]$.

Pooled Exponential QQ plot - here $\mathbf{X} \in \mathcal{D}$.



Goes back to to (?) Heffernan, J.E., Tawn, J.A. Extreme Value Analysis of a Large Designed Experiment: A Case Study in Bulk Carrier Safety. Extremes 4, 359–378 (2001)

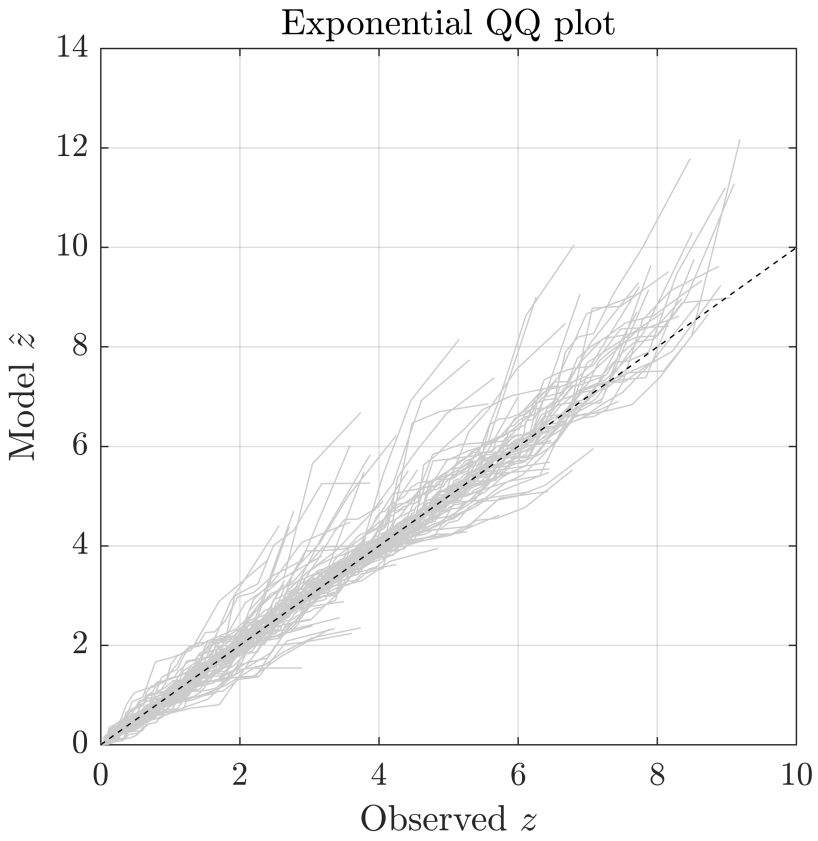
Limitation of Exponential QQ Plots

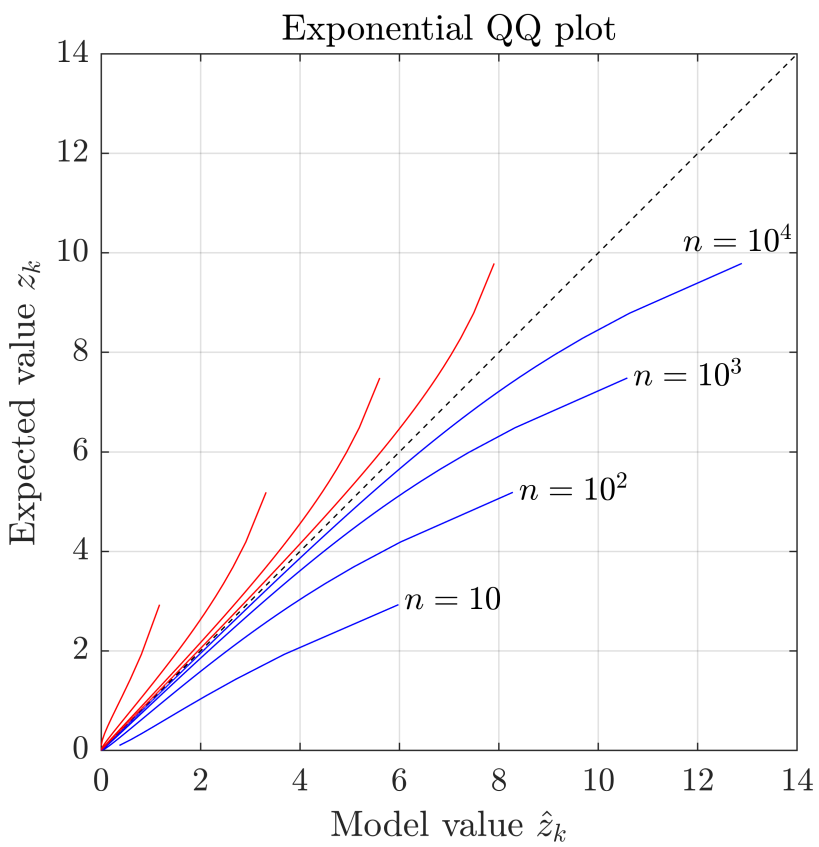
Compared to standard QQ plots, pooled Exponential QQ plots are independent of tail shapes and scales, and preserve information about the non-stationary distribution.

First idea: evaluate exponential QQ plot for each bin, $\mathcal{B}_1, \ldots, \mathcal{B}_B$, plot on same graph.

However, they are not invariant to sample size, and do not provide a visual assessment of the significance of differences between the model and observations.

Limitation of Exponential QQ Plots





To account for these effects, we need to consider the asymptotic sampling distributions of exceedance probabilities.

Since $Q_{(k)}$ is a random variable, and in many applications, the data generating distribution, $F_{Y|\boldsymbol{X}}$, is not known, we must make some 'best guess' of the sample probabilities.

It is well-known that $Q_{(k)} \sim \text{Beta}(k, n-k+1)$ (see e.g. David and Nagaraja (2003)). The expected values of the ranked exceedance probabilities are therefore $q_k := \mathbb{E}\left[Q_{(k)}\right] = k/(n+1)$.

Lots of literature creates pooled exponential QQ plots taking expectations on the exceedance probability scale, i.e., setting $z_k = -\log(k/(n+1))$. However, due to the nonlinear transformation $\mathbb{E}\left[Z_{(k)}\right] \neq -\log\left(\mathbb{E}\left[Q_{(k)}\right]\right)$.

A nice feature of the exponential distribution is that the expected values of the order statistics have a simple closed form expression David and Nagaraja (2003)

$$z_k := \mathbb{E}\left[Z_{(k)}\right] = H_n - H_{k-1}, \tag{1}$$

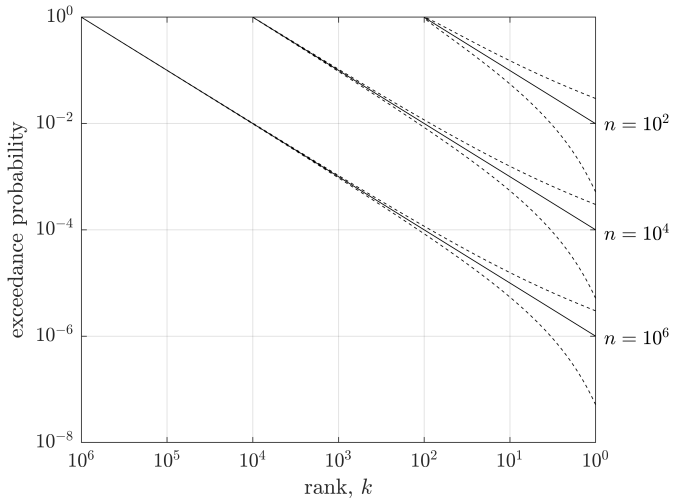
where $H_k = \sum_{j=1}^{k} \frac{1}{j}$ is the k-th harmonic number, and we define $H_0 = 0$ for convenience. Note that

$$H_n = \log(n) + \gamma + O(1/n), \tag{2}$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant. Therefore, the difference between z_1 and $-\log(q_1)$ tends to γ as $n \to \infty$. So, ignoring these differences can lead to biases in the diagnostics if results are aggregated over many local regions.

We opt to use the above expectation for plotting points, as we will create a standardised diagnostic with zero mean.

Asymptotic Result for Extremes



Theorem

For any fixed $k \in \mathbb{N}_{>0}$, the normalised exceedance probability $nQ_{(k)}$ of the k-th largest order statistic converges in distribution, as $n \to \infty$, to a gamma distributed variable, with shape parameter k and unit scale.

Asymptotic Result for Extremes

Corollary

For any fixed $k \in \mathbb{N}_{>0}$, the difference $D_k = \mathbb{E}\left[Z_{(k)}\right] - Z_{(k)}$ between the expected value of the k-th exponential order statistic and the observed value, converges in distribution as $n \to \infty$, to a log-gamma distributed variable, $D_{k,\infty}$, with shape parameter k, unit scale, and location $\mu_k = \gamma - H_{k-1}$, with density function

$$f_{D_{k,\infty}}(d) = \frac{1}{\Gamma(k)} \exp\left[k\left(d - \mu_k\right) - \exp\left(d - \mu_k\right)\right]. \tag{3}$$

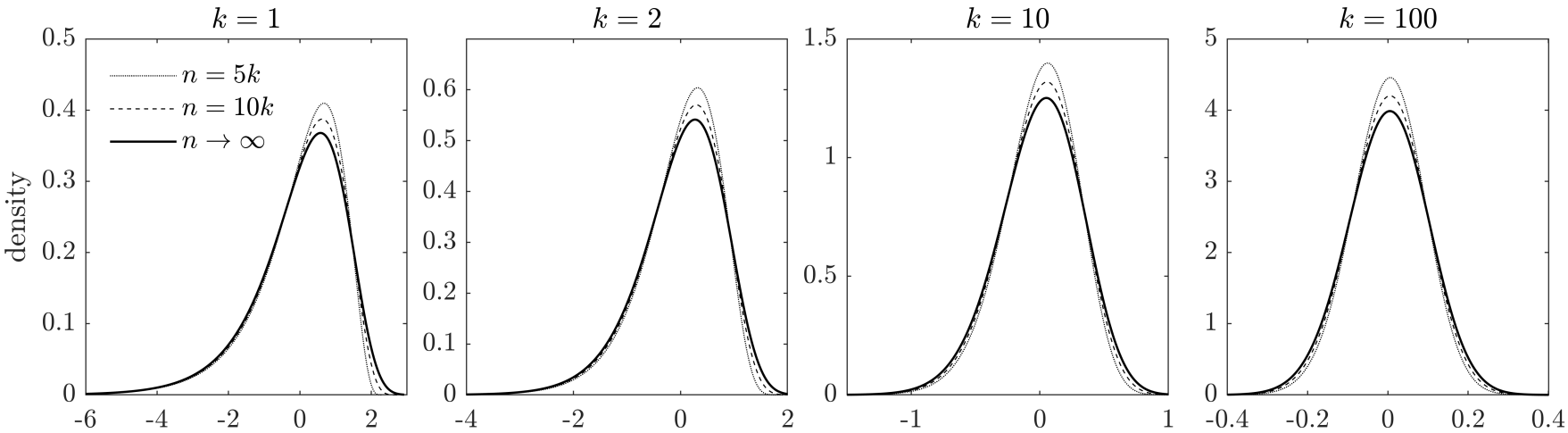


Figure: Densities of standardised exponential order statistics $D_k = \mathbb{E}\left[Z_{(k)}\right] - Z_{(k)}$, for various ranks k and sample sizes n.

Fact Convergence - **Implication**:

- Sampling distribution of extremes independent of n
- Enables rank-wise standardisation

Standardised Tail Plot

Define deviation

$$D_k = \mathbb{E}[Z_{(k)}] - \hat{Z}_{(k)}$$

Asymptotically ($n \to \infty$):

$$D_k \sim \text{log-Gamma}(\mu_k, k, 1),$$

with $\mu_k = \gamma - H_{k-1}$ and $\mathbb{E}_k = 0$.

- Plot D_k vs rank k
- Overlay theoretical confidence bands
- Highlights significant tail misfit
- Can get multiple realisations of D_k using local binning

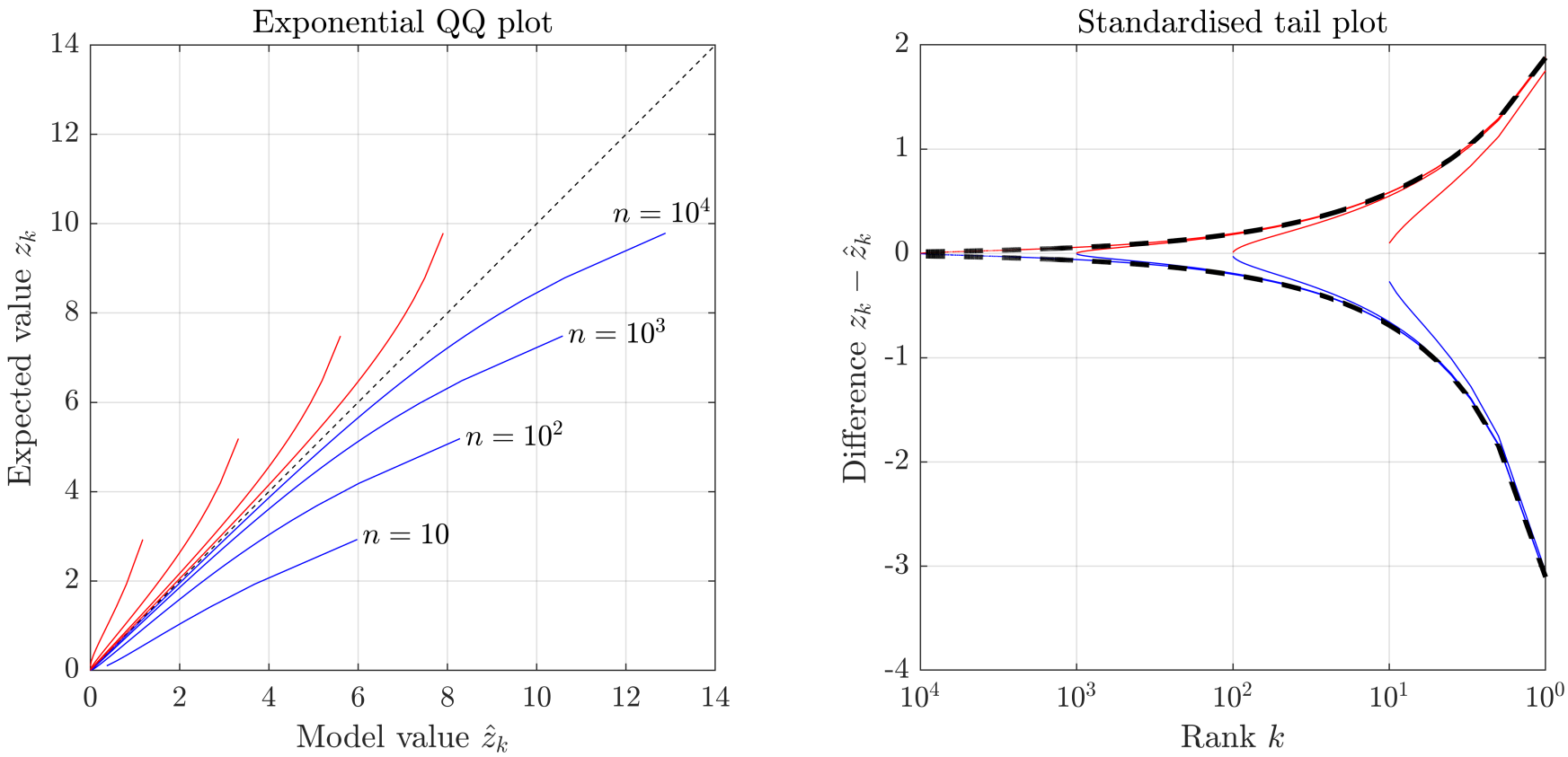


Figure: Top left: Two-sided 95% CI on exponential quantiles for sample sizes of $n = 10, 10^2, 10^3, 10^4$. Top right: Two-sided 95% CI for differences between model and expected exponential quantiles for the same sample sizes, together with 95% CI for the asymptotic distribution (black dashed lines).

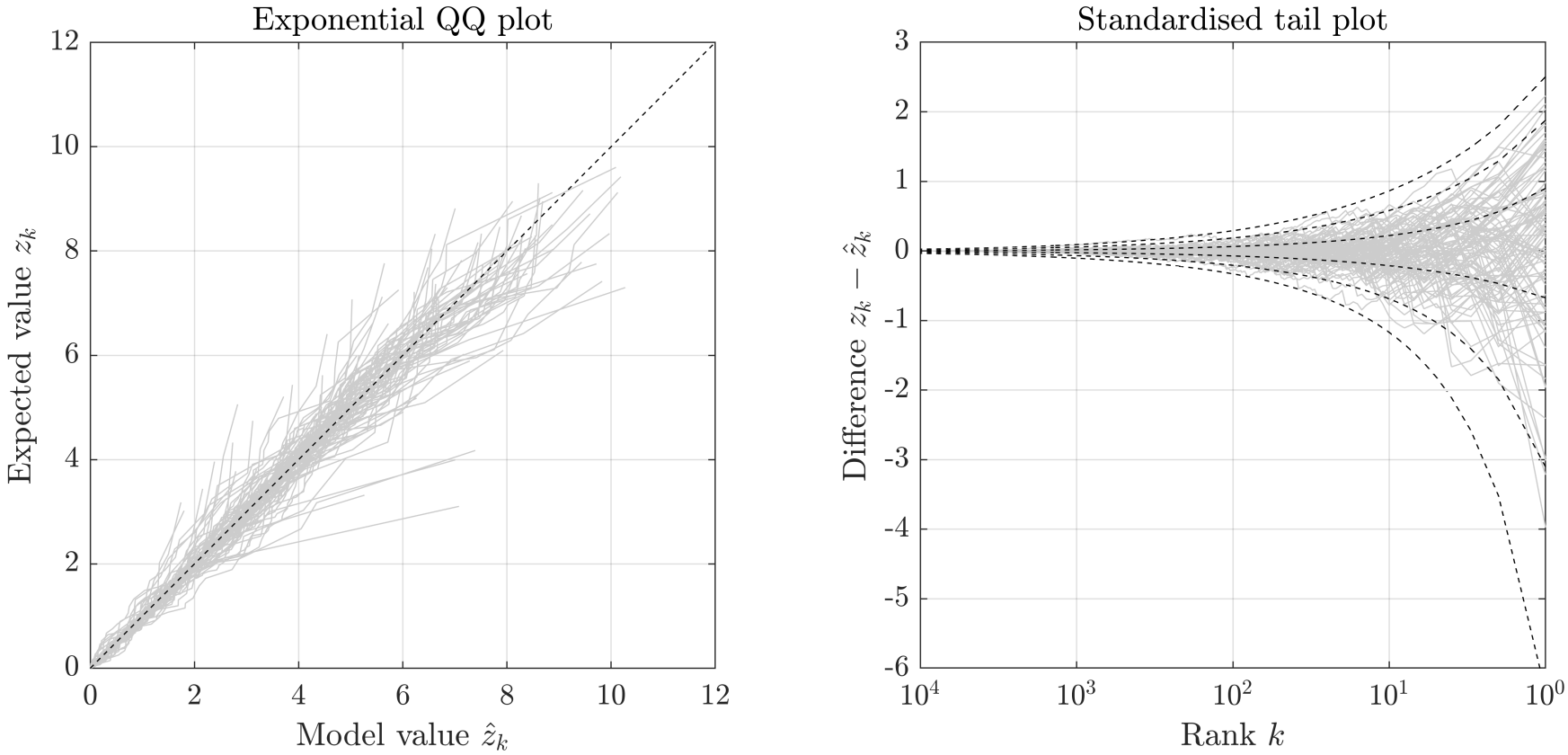


Figure: Lower left: Example simulations for 100 samples of random size N, where $\log_{10}(N) \sim \mathrm{U}(1, 4)$. Lower right: Same data transformed to standardised scale, together with quantiles of the asymptotic distribution at non-exceedance probabilities 0.001, 0.025, 0.25, 0.75, 0.975, and 0.999 (dashed lines).

Rank-independent Diagnostics

Exact finite-sample distribution:

$$Q_{(k)} \sim \text{Beta}(k, n - k + 1)$$

Transform to Gaussian scale:

$$b_k = \Phi^{-1}(\beta_k(\hat{q}_{(k)}))$$

Interpretation:

- $b_k \sim N(0, 1)$ under correct model
- Visualises significance across all ranks

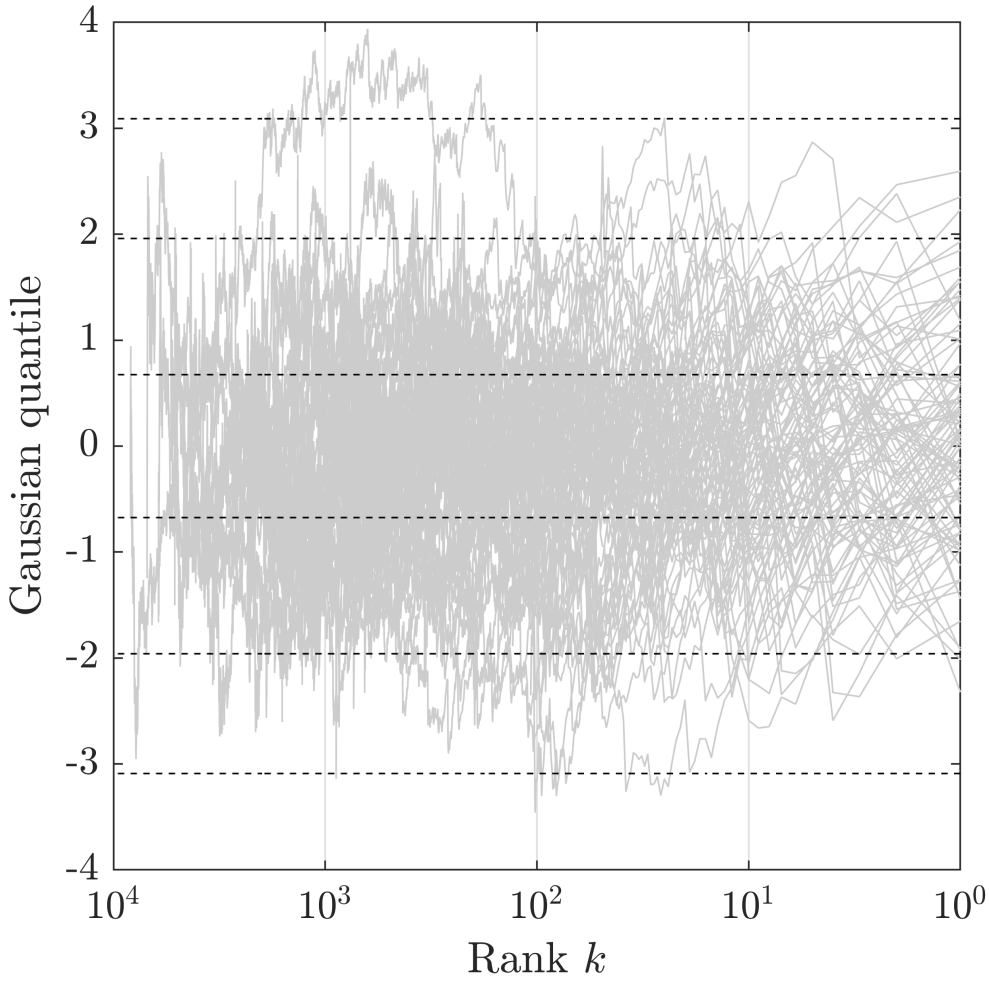


Figure: Transformed Gaussian quantiles. Each line grey corresponds to an individual sample. Dashed lines indicate normal quantiles at non-exceedance probabilities 0.001, 0.025, 0.25, 0.75, 0.975, and 0.999.

Global vs Local Diagnostics

- Global vs **Regional summaries**: goodness-of-fit per covariate region
- Visual diagnostics - pooled Exponential QQ plot, standardised tail plot, Gaussian plot, k-wise asymptotic sampling dist for standardised tail plot.
- Used in combination with metrics from Anderson–Darling (right hand; ADR) and Cramér–von Mises (CvM) tests

$$A^2 = n \int_0^1 (\hat{F}(y) - F(y))^2 w(y) \mathrm{d}y,$$

where

$$w(y) = [1 - F(y)]^{-1}, \quad w(y) = 1$$

Global summary Diagnostics

1. Estimate $\hat{F}_{Y|\mathbf{X}}$ using all $\mathbf{X} \in \mathcal{D}$.
2. Evaluate **global** integrated squared deviance

$$\sum_{k=1}^{n} D_k^2 = \sum_{k=1}^{n} (\hat{z}_{(k)} - z_{(k)})^2$$

3. Evaluate **global** ADR for $\hat{F}_{Y|\mathbf{X}}$, with p-value obtained via Monte Carlo.

Useful for model selection and hyperparameter tuning.

Regional summary Diagnostics

Split $\mathcal{D}$ into bins $\mathcal{B}_1, \ldots, \mathcal{B}_B$.

Then, for $b = 1, \ldots, B$:

1. Evaluate **regional** ADR_b for $\hat{F}_{Y|\mathbf{X}\in\mathcal{B}_b}$
2. Evaluate $\{\hat{q}^{(b)}_{(k)}\}_{k=1,\ldots,n_b}$.

Then

1. CvM test for uniformity of p-values for $\{\text{ADR}_b\}_{b=1,\ldots,B}$
2. For $k = 1, \ldots, K$, CvM test for uniformity for $\{\beta(\hat{q}^{(b)}_{(k)})\}_{b=1,\ldots,B}$

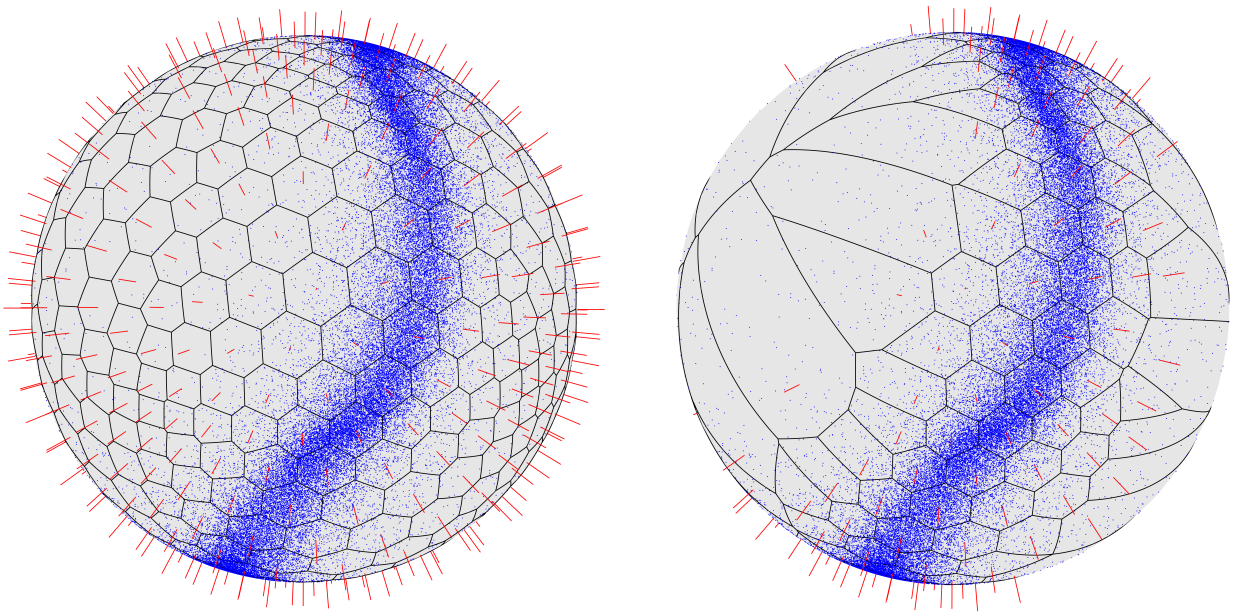
Useful for model selection and hyperparameter tuning.

Example - Deep SPAR Model

- $d = 5$ Gaussian copula for $\mathbf{X}$, $n = 20000$, Laplace margins.
- Multivariate extremes via pseudo-polar coordinates
- Conditional radial component $R \mid \|\mathbf{X}\| = \mathbf{w}$

$$f_{R|\|\mathbf{x}\|}(r|w) \propto r^{\beta(w)-1} \exp\left(-\frac{r}{\sigma(w)}\right).$$

Diagnostics applied across 360 hypersphere partitions.



Threshold Selection via Diagnostics

- 5000 candidate models fitted - Deep exponential regression model above some high τ-quantile.
- First requires threshold model with threshold non-exceedance probability τ
- Choose τ maximising regional ADR p-values

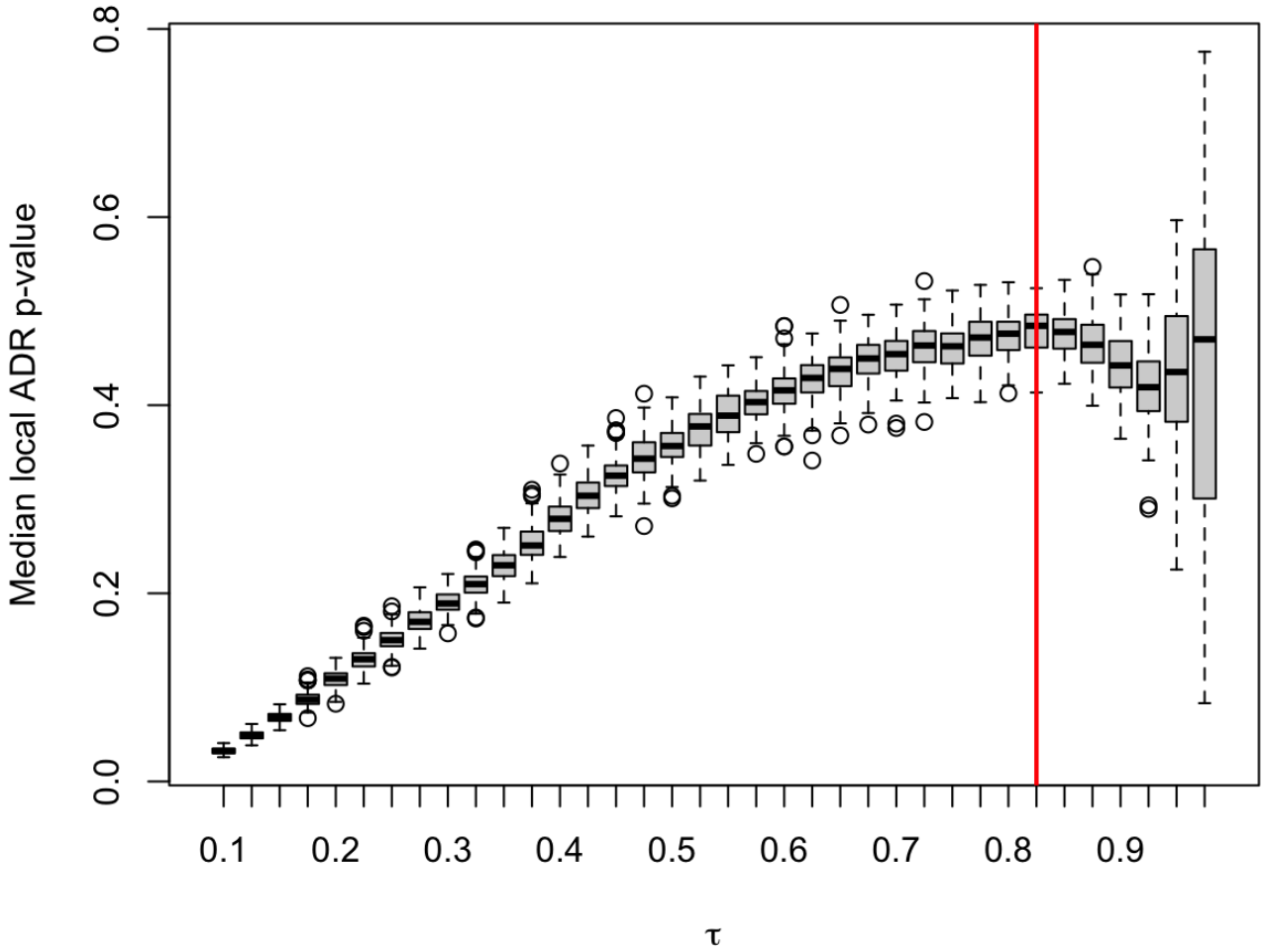
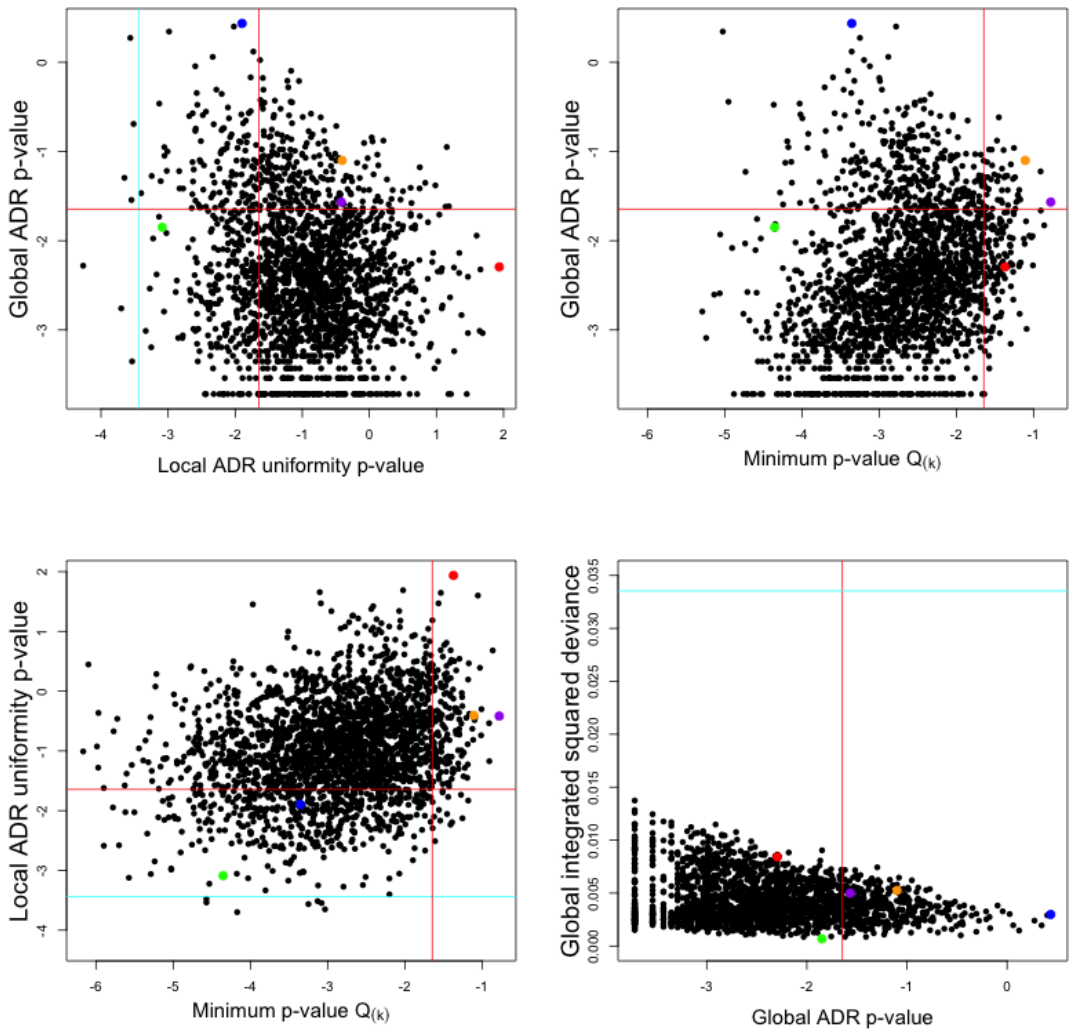


Figure: Box-plots of the empirical median of local ADR p-values pooled across 5000 estimated candidate model. The x-axis gives the candidate non-exceedance probability τ, and the red vertical line passes through $\tau = 0.825$.

Local diagnostics

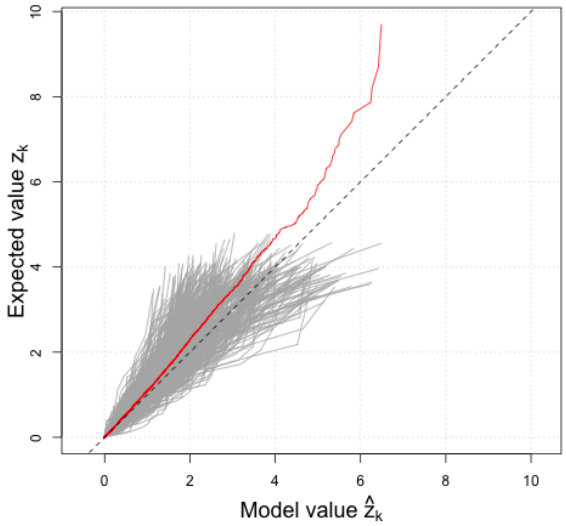
- A further 2500 candidate models fitted with $\tau = 0.825$.
- Collection of global and local diagnostics/summary statistics evaluated for each model
- Each bin has 10–70 **test** points.

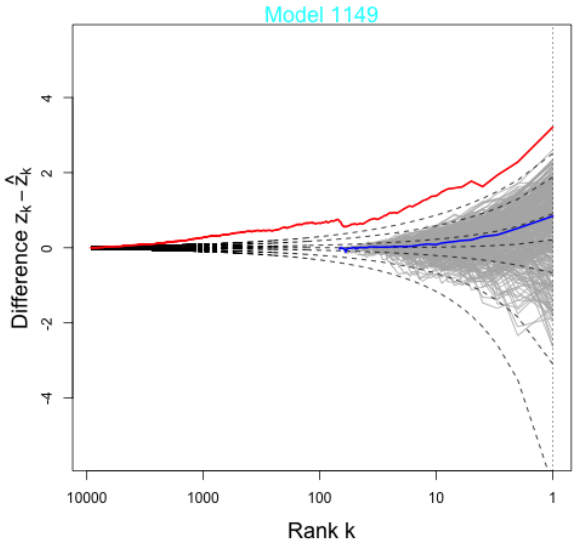
Local diagnostics - scatter

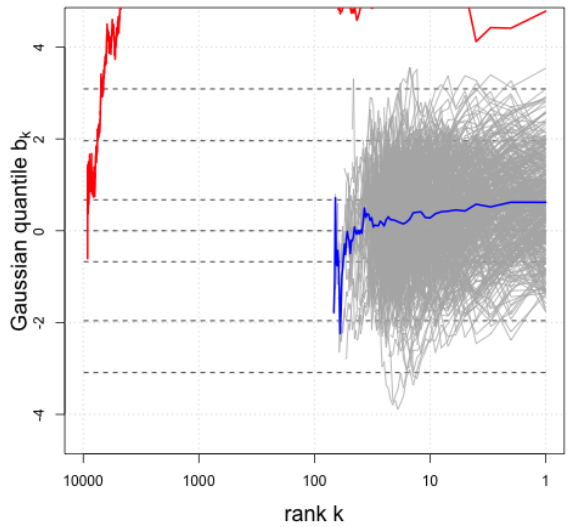


Local diagnostics - cyan

Fails all tests!

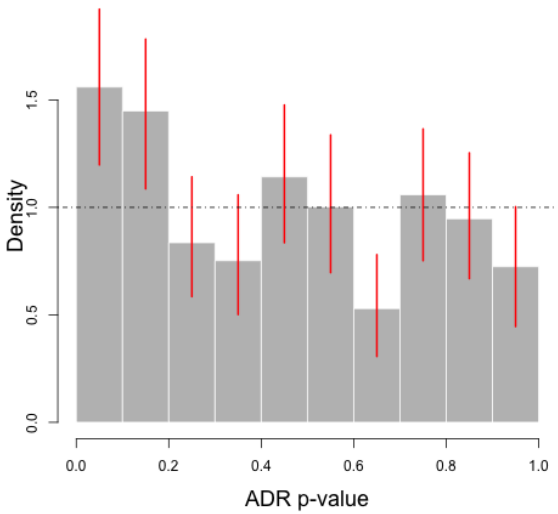


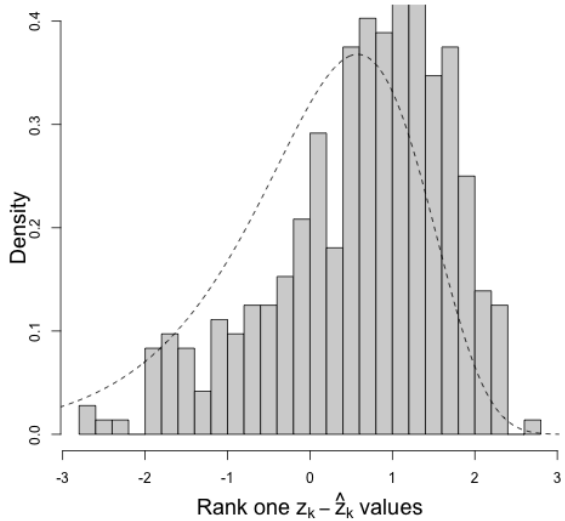


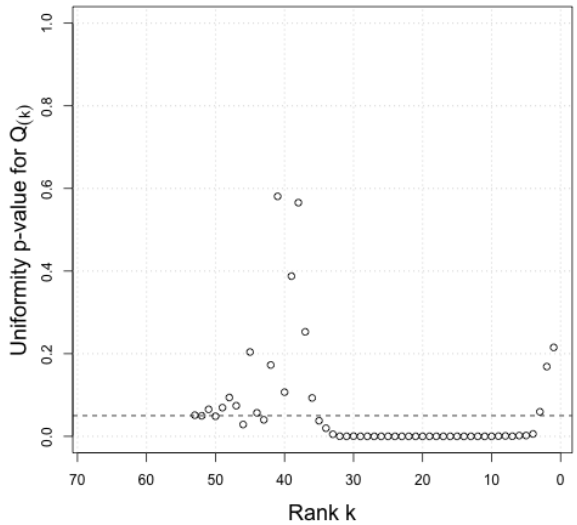


Local diagnostics - cyan

Fails all tests!

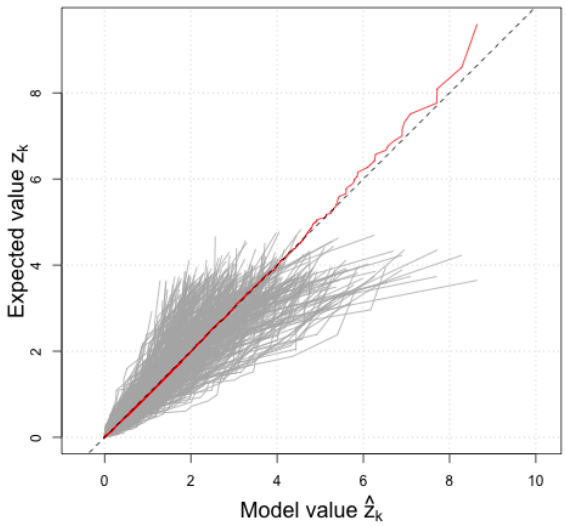


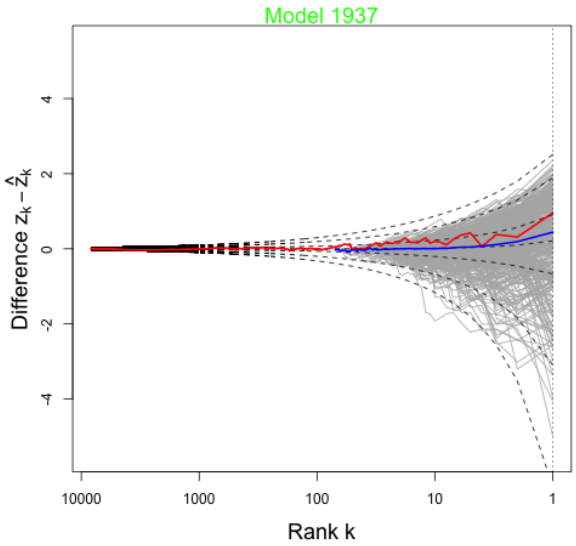


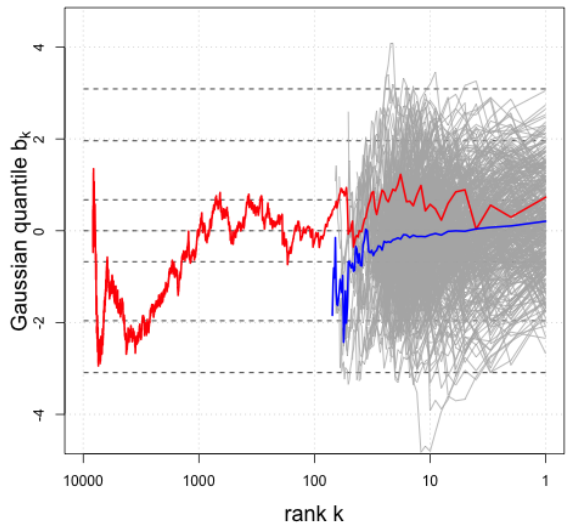


Local diagnostics - green

Good global fit, poor local fits!

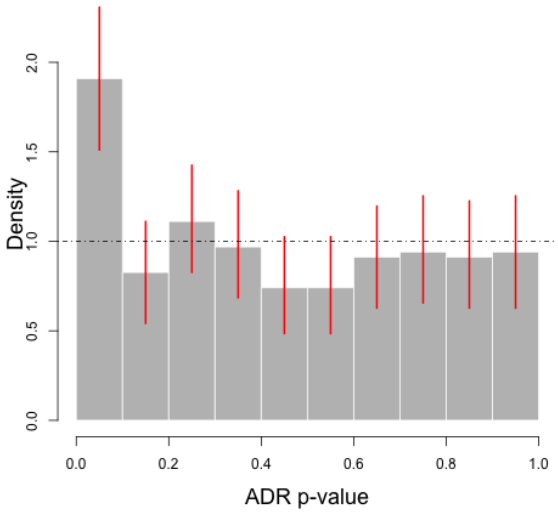


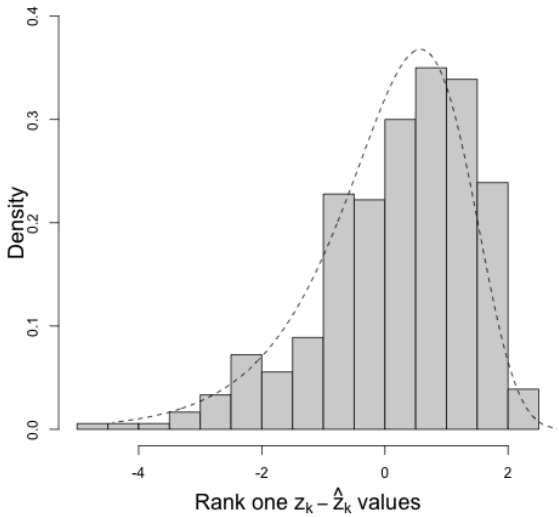


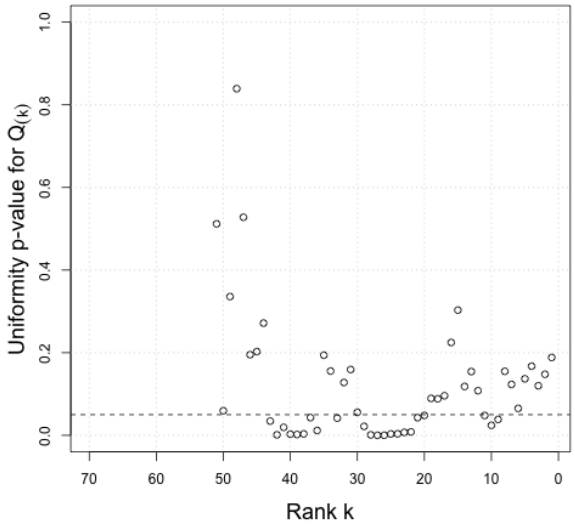


Local diagnostics - green

Good global fit, poor local fits!

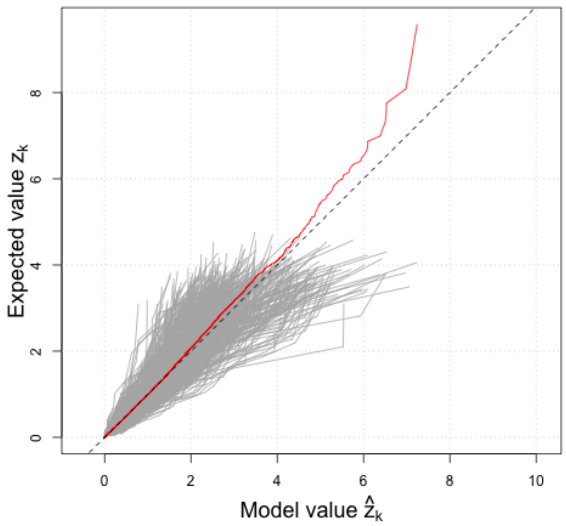


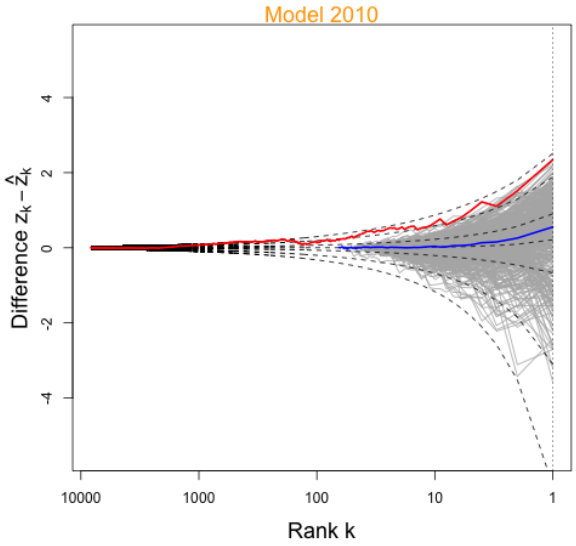


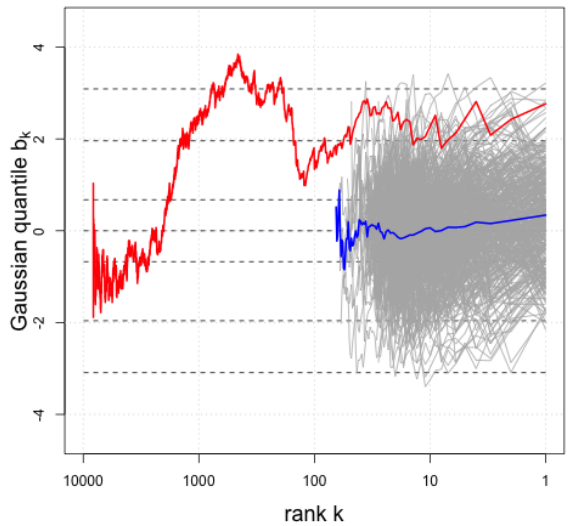


Local diagnostics - purple

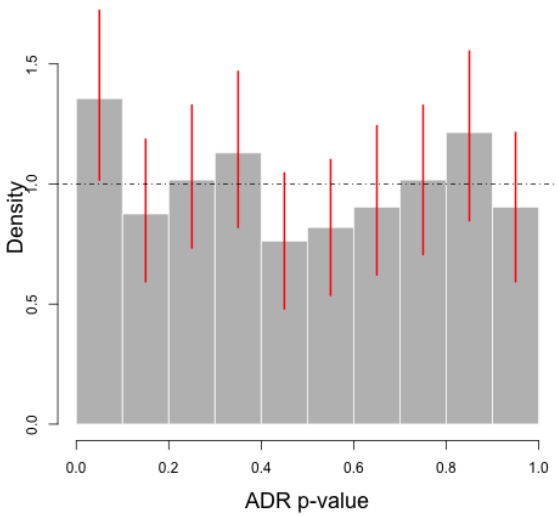
Best of both worlds

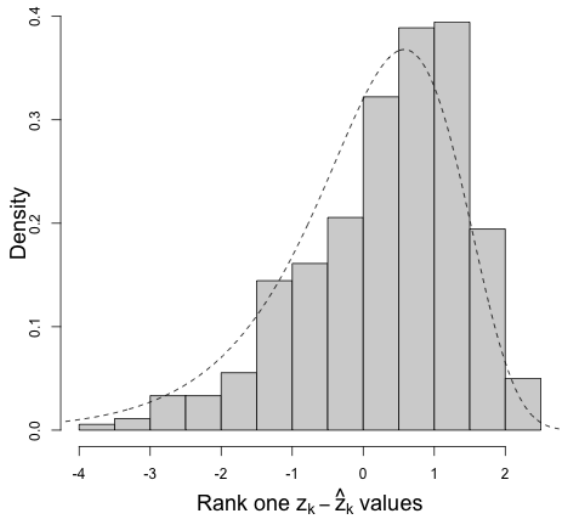


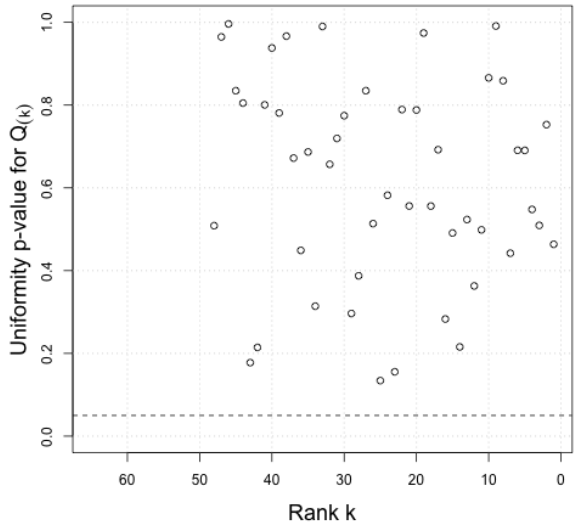




Best of both worlds Good global fit, poor local fits!

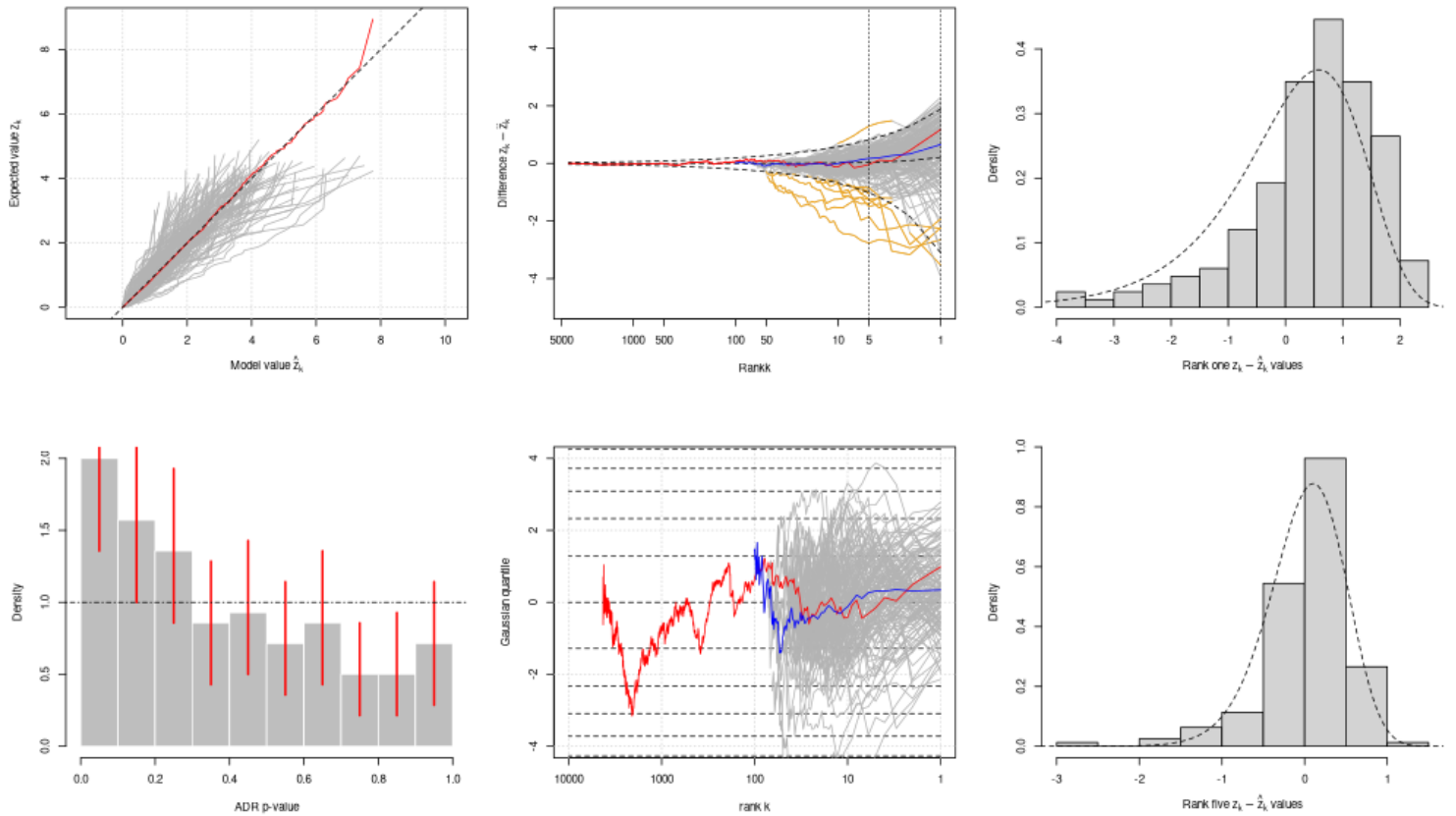






Actionable insights?

Different application: i) data from surrogate wave model; ii) deep GPD regression model with discrete set $\mathcal{D} \subset [0,1]^3$.



References I

David, H. and Nagaraja, H. (2003). *Order Statistics*. Wiley Series in Probability and Statistics, third edition.