Modelling non-stationarity in asymptotically independent extremes

Callum Murphy-Barltrop, Jenny Wadsworth
Lancaster University, UK

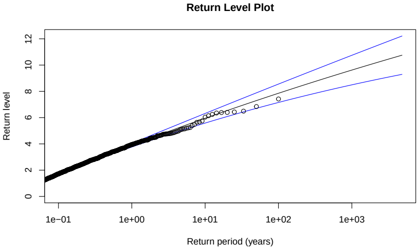
Energy Forecasting Innovation Conference, 24 May 2022

Callum's work:

- PhD project (2019-2023) with the Office for Nuclear Regulation (ONR) via the STOR-i centre for doctoral training
- **Main theme**: investigating use of multivariate extreme value theory in the context of nuclear regulation
- Follows scoping project (Simpson, 2017) with LU seed money to investigate concept of a "multivariate hazard curve"
- Focus on "return curves" as a way to summarize risk associated to combined effect of two (or potentially more) variables
- Particular challenges:
  - Estimation methodology under realistic dependence scenarios
  - **Non-stationarity of environmental variables**

Hazard curves and return curves

A **hazard curve** (often called a **return level plot** by EVA practitioners) depicts levels associated to a certain **annual exceedance probability**, $1/N$.

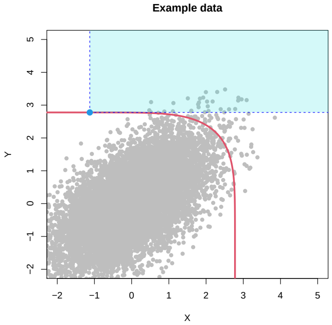


- Mathematically: plot of points $\{(N, x_N)\}$ for N and x_N satisfying
$$\Pr(X > x_N) = 1/(n \times N),$$
where n is the number of observations of X per year
- x_N often called N-year return level. N is the return period.
- Summarizes the tail of a (stationary) univariate distribution

Hazard curves and return curves

A **return curve** provides a possible extension of the definition. Now consider the N-year return curve:

$$\mathrm{RC}(N) = \left\{ (x_N, y_N) : \Pr(X > x_N, Y > y_N) = \frac{1}{n \times N} \right\}$$



- Because the set of points $\{(x_N, y_N)\}$ already represents a curve, helpful to plot separately for different N
- Summarizes the tail of a (stationary) bivariate distribution

- Nuclear facilities should be designed to withstand extreme events with a return period of $N = 10,000$ years
  - ONR Safety Assessment Principles: values should be derived conservatively to take account of data and model uncertainties
- What if the 10,000 year event changes over time?
- Conservative principles suggest design to withstand the worst such event over the time period
- Desire to understand more about what this means in the context of more than one hazard
- **This talk**: estimation of return curves for **non-stationary bivariate distributions**

Non-stationary return curves

Assume observations $(X_t, Y_t), t = 1, \ldots, T$, whose distribution is influenced by covariates $\boldsymbol{Z}_t$.

Non-stationary N-year return curve:

$$\mathrm{RC}_{\boldsymbol{z}_t}(N) = \left\{ (x_N, y_N) : \Pr(X_t > x_N, Y_t > y_N | \boldsymbol{Z}_t = \boldsymbol{z}_t) = \frac{1}{n \times N} \right\}$$

Two sources of non-stationarity:

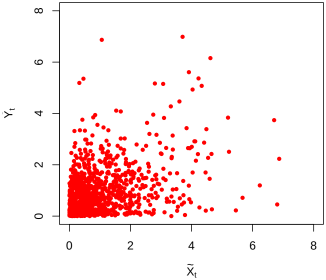
- **Marginal distributions** $X_t \sim F_{X_t}$, $Y_t \sim F_{Y_t}$
  - Well studied - but note we need to estimate full non-stationary cdf, not just the tail
- **Dependence structure / copula**
  - Sparse literature, especially for extremes

Non-stationary dependence

Joint distribution of $(F_{X_t}(X_t), F_{Y_t}(Y_t))$ called the **copula**:

$$C_t(u, v) = \Pr(F_{X_t}(X_t) \leq u, F_{Y_t}(Y_t) \leq v)$$

Assume this changes with covariates $\boldsymbol{Z}_t$.



Non-stationary extremal dependence measures

Extremal dependence often summarized by measures derived from the copula:

Tail dependence coefficient

$$\chi_t = \lim_{u \to 1} \frac{\Pr(F_{X_t}(X_t) > u, F_{Y_t}(Y_t) > u)}{1 - u} \in [0, 1]$$

Useful if variables exhibit so-called **asymptotic dependence**, meaning $\chi_t > 0$. Otherwise $\chi_t = 0$ even as the copula changes with $\boldsymbol{Z}_t$.

Extremal dependence often summarized by measures derived from the copula:

Tail dependence coefficient

$$\chi_t = \lim_{u \to 1} \frac{\Pr(F_{X_t}(X_t) > u, F_{Y_t}(Y_t) > u)}{1 - u} \in [0, 1]$$

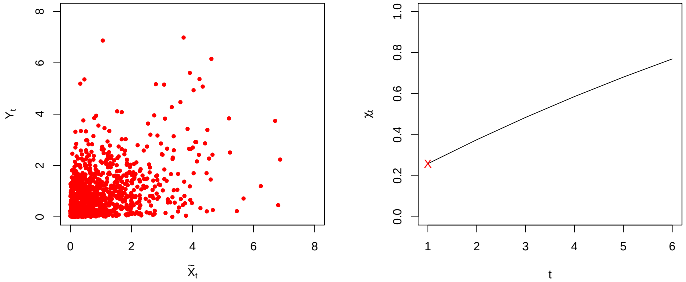
Useful if variables exhibit so-called **asymptotic dependence**, meaning $\chi_t > 0$. Otherwise $\chi_t = 0$ even as the copula changes with $\boldsymbol{Z}_t$.

Residual tail dependence coefficient

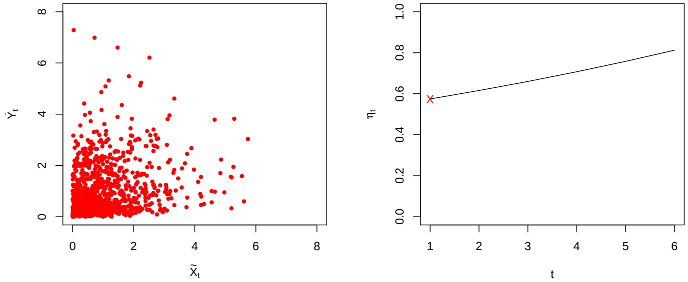
$$\eta_t = \lim_{u \to 1} \frac{\log(1 - u)}{\log \Pr(F_{X_t}(X_t) > u, F_{Y_t}(Y_t) > u)} \in (0, 1]$$

Useful if variables exhibit **asymptotic independence**, where typically $\eta_t < 1$. Otherwise $\eta_t = 1$ even as the copula changes with $\boldsymbol{Z}_t$.

Non-stationary extremal dependence measures



Non-stationary extremal dependence measures



Returning to return curves

- Measures like η and χ summarize dependence in the region where both variables are large
- Do not provide enough information to construct a return curve

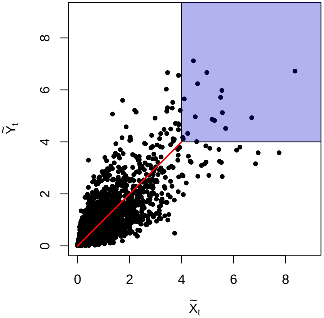
Focus on the **angular dependence function** (Wadsworth & Tawn, 2013) which describes extremal dependence in any direction where at least one variable is large.

Angular dependence function

Let $\tilde{X}_t = -\log(1 - F_{X_t}(X_t))$, $\tilde{Y}_t = -\log(1 - F_{Y_t}(Y_t))$ have stationary exponential margins.

Residual tail dependence coefficient

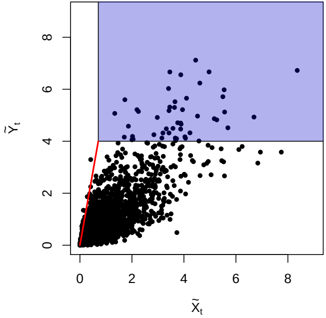
$$\Pr\left(\min\left\{\tilde{X}_t, \tilde{Y}_t\right\} > x + u \middle| \min\left\{\tilde{X}_t, \tilde{Y}_t\right\} > u\right) \to \exp\left\{-x/\eta_t\right\}, \qquad u \to \infty$$



Let $\tilde{X}_t = -\log(1 - F_{X_t}(X_t))$, $\tilde{Y}_t = -\log(1 - F_{Y_t}(Y_t))$ have stationary exponential margins.

Angular dependence function For $w \in (0, 1)$

$$\Pr\left(\min\left\{\frac{\tilde{X}_t}{w}, \frac{\tilde{Y}_t}{1-w}\right\} > x + u \,\middle|\, \min\left\{\frac{\tilde{X}_t}{w}, \frac{\tilde{Y}_t}{1-w}\right\} > u\right) \to \exp\{-x\lambda_t(w)\}$$

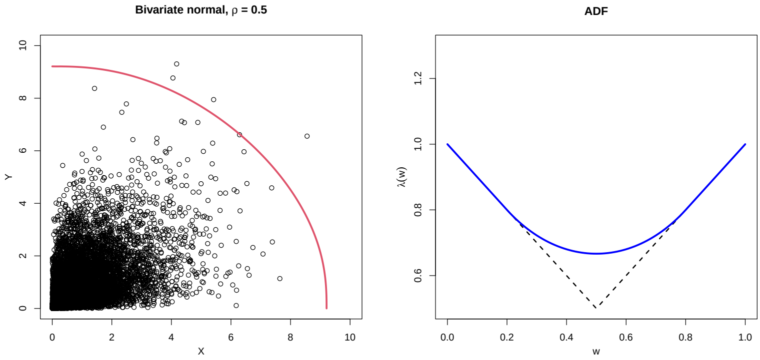


- Generalizes concept of residual tail dependence coefficient: $1/\eta = 2\lambda(1/2)$
- $\lambda(w) \geq \max(w, 1-w)$
- Asymptotic dependence: $\lambda(w) = \max(w, 1-w)$
- Independence: $\lambda(w) = 1$

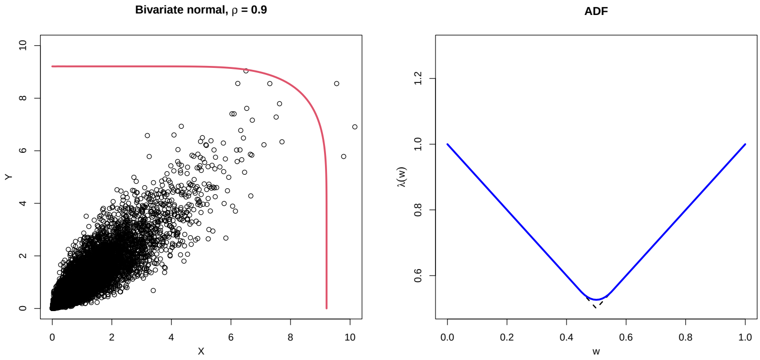
Can be considered in a non-stationary context (like η_t, χ_t). Denote this by $\lambda_t(w)$.

Issue with all non-stationary dependence measures is **estimation**: we typically do not have repeated samples for a given covariate value $\boldsymbol{z}_t$.

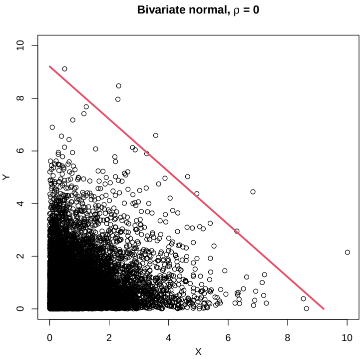
Examples

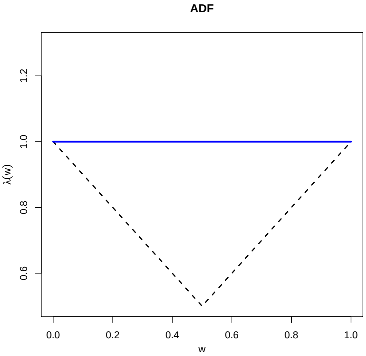


Examples

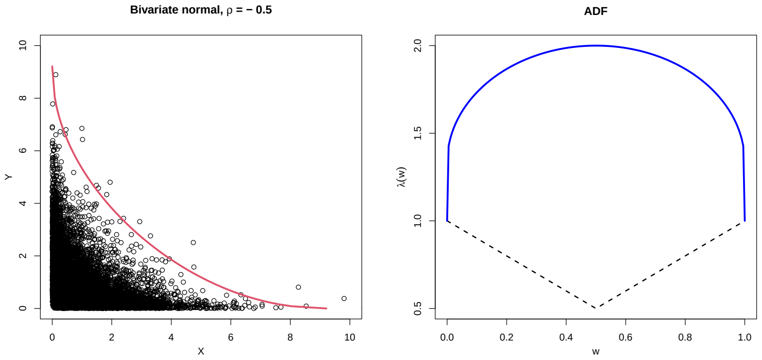


Examples





Examples


Bivariate normal, ρ = − 0.5
Y
X
0
2
4
6
8
10
ADF
λ(w)
w
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

Define $K_{w,t} := \min\left\{\frac{\tilde{X}_t}{w}, \frac{\tilde{Y}_t}{1-w}\right\}$. For a high threshold u we have

$$\Pr(K_{w,t} > v + u | K_{w,t} > u, \boldsymbol{Z}_t = \boldsymbol{z}_t) \approx \exp\left\{-\lambda_t(w)v\right\}$$

For fixed w consider quantile regression on the distribution of $K_{w,t}|\boldsymbol{Z}_t = \boldsymbol{z}_t$. For $q_1 < q_2$ near 1, we find sequences $u_{w,t}$, $v_{w,t}$ such that

$$\Pr(K_{w,t} > u_{w,t}|\boldsymbol{Z}_t = \boldsymbol{z}_t) = 1 - q_1$$
$$\Pr(K_{w,t} > v_{w,t} + u_{w,t}|\boldsymbol{Z}_t = \boldsymbol{z}_t) = 1 - q_2$$

Combining expressions:

$$\frac{1 - q_2}{1 - q_1} \approx \exp\left\{-\lambda_t(w)v_{w,t}\right\}$$

Initial estimator at point w

$$\hat{\lambda}_t(w) = -\frac{1}{v_{w,t}} \log\left(\frac{1-q_2}{1-q_1}\right)$$

For stability, we repeat this procedure over several pairs of high quantiles $(q_{1,j}, q_{2,j}), j = 1, \ldots, m$ and take the average as an estimator:

$$\hat{\lambda}_t(w) \approx -\frac{1}{m}\sum_{j=1}^{m} \frac{1}{v_{w,t,j}} \log\left(\frac{1-q_{2,j}}{1-q_{1,j}}\right)$$

Smoothing over w

Procedure is repeated on a grid of w values $\{w_1, \ldots, w_k\} \in (0, 1)$ giving

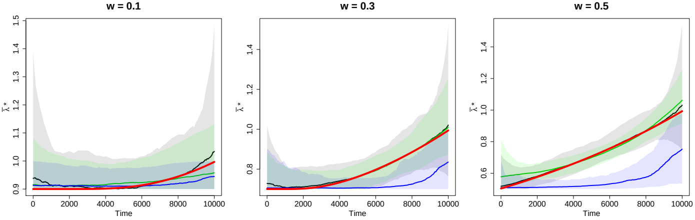
$$\hat{\lambda}_t(w_1), \ldots, \hat{\lambda}_t(w_k)$$

for each value of our covariates $\boldsymbol{z}_t$.

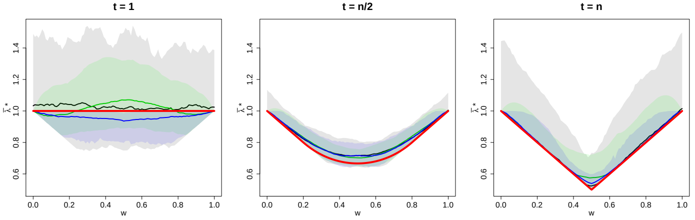
To smooth over w, we use **Bernstein polynomials**, with the non-smoothed estimates as input into a suitable objective function.

Performance of estimators

Performance over covariate (time):



Performance across $w \in (0, 1)$ at three fixed time points:



Deriving a return curve from $\lambda_t(w)$

Let $p = 1/(n \times N)$, and for any $w \in (0, 1)$, define $\{r_{w,t}\}$ as

$$r_{w,t} := -\frac{1}{\hat{\lambda}_t(w)} \log\left(\frac{p}{1-q_1}\right),$$

implying $\frac{p}{1-q_1} = \exp\{-r_{w,t}\hat{\lambda}_t(w)\}$.

Deriving a return curve from $\lambda_t(w)$

Let $p = 1/(n \times N)$, and for any $w \in (0,1)$, define $\{r_{w,t}\}$ as

$$r_{w,t} := -\frac{1}{\hat{\lambda}_t(w)} \log\left(\frac{p}{1-q_1}\right),$$

implying $\frac{p}{1-q_1} = \exp\{-r_{w,t}\hat{\lambda}_t(w)\}$.

Define $(x_{w,t}, y_{w,t}) := (w(r_{w,t} + u_{w,t}), (1-w)(r_{w,t} + u_{w,t}))$. We have

$$\begin{aligned}
\Pr(X_t > x_{w,t}, Y_t > y_{w,t} \mid \mathbf{Z}_t = \mathbf{z}_t) &= \Pr(K_{w,t} > r_{w,t} + u_{w,t} \mid \mathbf{Z}_t = \mathbf{z}_t) \\
&= \Pr(K_{w,t} > r_{w,t} + u_{w,t} \mid K_{w,t} > u_{w,t}, \mathbf{Z}_t = \mathbf{z}_t) \\
&\quad \times \Pr(K_{w,t} > u_{w,t} \mid \mathbf{Z}_t = \mathbf{z}_t) \\
&\approx \exp\{-r_{w,t}\hat{\lambda}_t(w)\} \Pr(K_{w,t} > u_{w,t} \mid \mathbf{Z}_t = \mathbf{z}_t) \\
&= \frac{p}{1-q_1} \times 1 - q_1 = p,
\end{aligned}$$

meaning that the set $\{(x_{w,t}, y_{w,t})\}$ over a fine grid of points $\mathcal{W} = \{w_1, \ldots, w_k\}$ provides an approximation of $\mathrm{RC}_{\mathbf{z}_t}(N)$ on exponential margins.

Again, improve stability of estimation via **averaging over quantiles** $q_{1,j}$, $j = 1, \dots, m$.

Final estimator of the return curve on exponential margins:

$$\overline{\mathrm{RC}}_{\mathbf{z}_t}(N) = \left\{ \left(\sum_{j=1}^{m} x_{w,t}^{j}/m, \sum_{j=1}^{m} y_{w,t}^{j}/m \right) \right\}_{w \in \mathcal{W}}.$$

Last step: reverse marginal transformation to **original margins**. For estimates $\hat{F}_{X_t}, \hat{F}_{Y_t}$ of the marginal cdfs, apply transformation

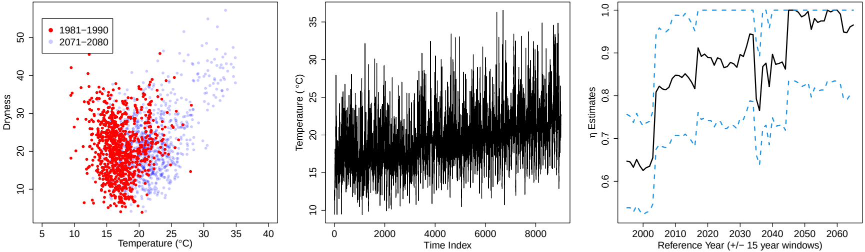
$$\left(\hat{F}_{X_t}^{-1}(1 - e^{-x}), \hat{F}_{Y_t}^{-1}(1 - e^{-y}) \right)$$

to the coordinates of the return curve.

Motivating example - UKCP18 data

- We consider 1980-2080 **temperature** and **relative humidity** projections at a location near to an existing licensed site
- RCP8.5 “worst-case” emissions scenario
- Focus on **summer data** as trends in dependence vary by season
- Let $0 < \mathrm{RH} < 100$ represent relative humidity. We define a **dryness** variable: $\mathrm{Dr} := 100 - \mathrm{RH}$
- **Combination** of high temperature and high dryness relevant for **nuclear safety** (Knochenhauer and Louko, 2004)
- Relevant covariate is **time**: $\boldsymbol{z_t} = t$ (true drivers for change are incorporated into climate model)

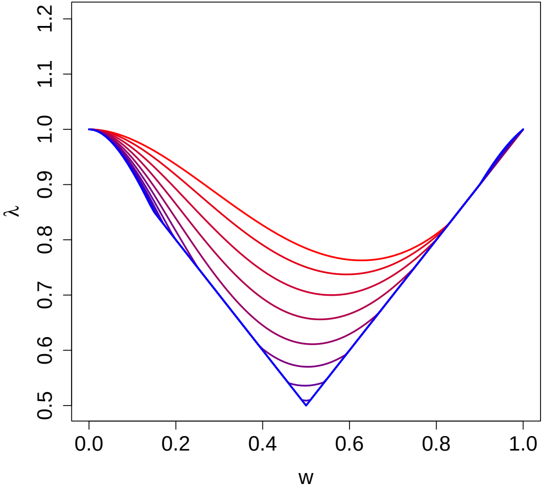
Trends in the data



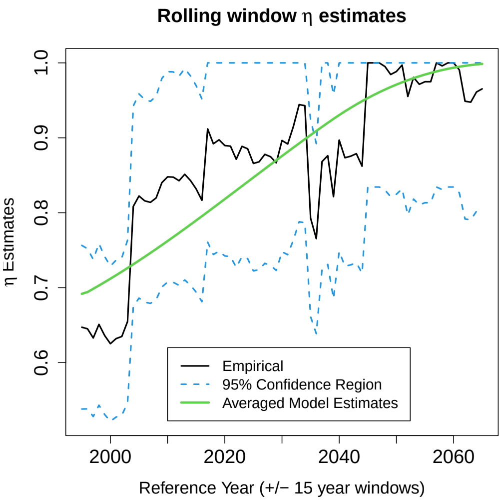
Modelling procedure

1. **Remove marginal trends** using methods proposed in Davison & Smith (1990) and Eastoe & Tawn (2009)
   1.1 Assume location-scale model: $X_t = \mu_{X,t} + \sigma_{X,t} R_t$, $Y_t = \mu_{Y,t} + \sigma_{Y,t} S_t$
   1.2 Estimate $\mu_{X,t}, \mu_{Y,t}, \sigma_{X,t}, \sigma_{Y,t}$ using GAMs
   1.3 Derived variables (R_t, S_t) should be close to stationary
   1.4 Fit non-stationary generalized Pareto distribution to tails to capture any residual non-stationarity in the tails
2. **Transform** data to exponential margins
3. Obtain **estimate** of **non-stationary ADF** through quantile regression + smoothing
4. **Calculate** return curve estimates up to the **year 2080**
5. **Back-transform** estimates to original margins

Non-stationary ADF

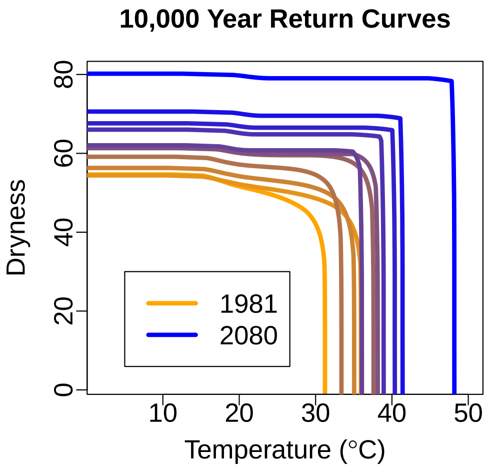


η_t estimates



Model estimate given by $\hat{\eta}_t = [2\hat{\lambda}_t(0.5)]^{-1}$

Estimates of $\mathrm{RC}_{\boldsymbol{z}_t}(10,000)$:



Conclusions

- Estimated return curves demonstrate **increasing marginal trends** for temperature and dryness, and **increasing dependence** over time
- Severe high temperature-dryness combinations **much more likely** in future climate scenarios (based on RCP8.5 emissions)
- Estimates suggest change from asymptotic independence to asymptotic dependence by end of timeframe

Discussion

- Framework is best suited to asymptotically independent variables: under asymptotic dependence, $\lambda_t(w) = \max(w, 1 - w)$ does not depend on $\boldsymbol{z}_t$
- But return curve estimates do inherit non-stationarity as they depend on estimates of the sequence $\{u_{w,t}\}$
- Reassuring for data like these that appear to transition between states
- Other methodology may be preferable for purely asymptotically dependent data (e.g. Castro-Camilo et al., 2018)
- **Uncertainty estimation**: Difficult for non-stationary data. Current approach: (block) bootstrap data in 5 year intervals, treating these as stationary

Thanks for your attention!

Murphy-Barltrop, C.J.R. and Wadsworth, J.L. (2022)
Modelling non-stationarity in asymptotically independent extremes *arXiv:2203.05860*

Castro-Camilo, D., de Carvalho, M. and Wadsworth, J.L. (2018)
Time-varying extreme value dependence with applications to leading European stock markets *Annals of Applied Statistics*

Davison, A.C. and Smith, R.L. (1990)
Models for exceedances over high thresholds *JRSSB*

Eastoe E.F. and Tawn, J.A. (2009)
Modelling non-stationary extremes with application to surface level ozone *JRSSC*

Knochenhauer, M. and Louko, P. (2004)
Guidance for External Events Analysis
In *Probabilistic Safety Assessment and Management*, pages 1498–1503. Springer, London.

Simpson, E.S. (2017)
Introduction to Extreme Value Theory and Constructing Hazard Curves
Available at `https://www.onr.org.uk/documents/2017/onr-rrr-054.pdf`

Wadsworth, J.L. and Tawn, J.A. (2013)
A new representation for multivariate tail probabilities *Bernoulli*