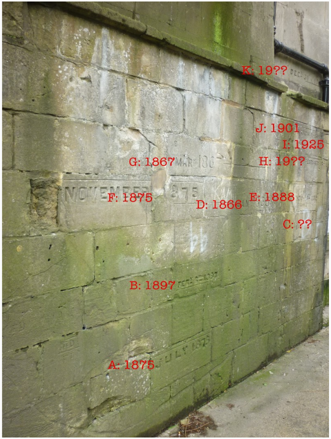
A clustering framework for conditional extremes models

Environmental and Ecological Statistics Conference 2025 Talk,
1st June, 2025,

Paddy O'Toole, University of Bath
Supervised by Christian Rohrbeck and Jordan Richards (University of Edinburgh)

Introduction


K: 19??
J: 1901
I: 1925
G: 1867
H: 19??
F: 1875
D: 1866
E: 1888
C: ??
B: 1897
A: 1875

Problem

- Often want to estimate

$$\mathbb{P}(X > x, Y > y) = \mathbb{P}(Y > y \mid X > x)\mathbb{P}(X > x)$$

  for large x, y
- "concomitant"/concurrent extreme events for random vector $\mathbf{X}$ often particularly devastating
- Goal: **identify trends by clustering sites with similar tail dependence**

Dependence Modelling

Asymptotic dependence

- Coefficient of extremal dependence $\chi \in [0, 1]$,

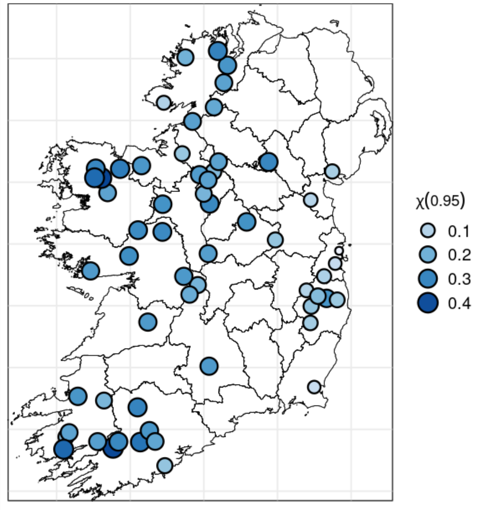
$$\chi = \lim_{u \to 1} \mathbb{P}\big[F_1(X_1) > u \mid F_2(X_2) > u\big]$$

- (Increasingly strong) asymptotic dependence for $\chi > 0$.
- However, χ only gives summary; inference requires **dependence model**.

Coles, J. Heffernan, and J. Tawn, *"Dependence measures for extreme value analyses"* (1999)

Ireland

- Precipitation[1] & wind speed[2] data for 59 sites across Ireland, Winter months (Oct-Mar) 1990-2020

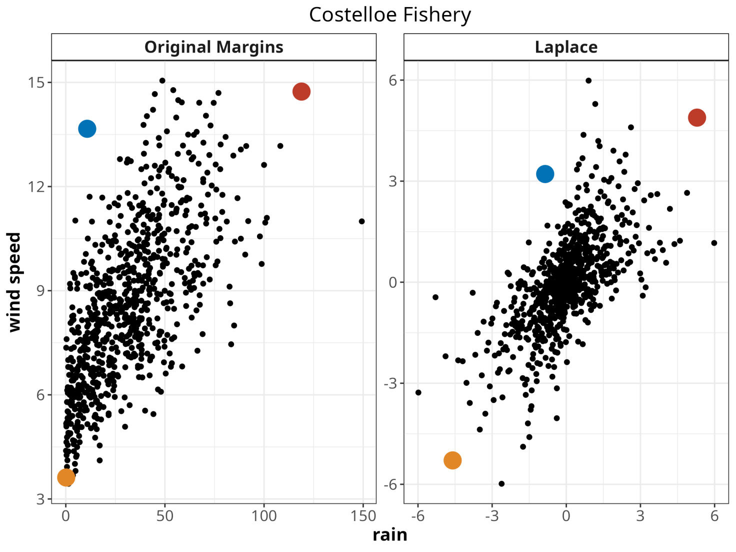


[1] Met Éireann weekly aggregate

[2] ERA5 reanalysis weekly mean of daily maxima

Conditional extremes

Marginal transformation



Conditional extremes

- Heteroskedastic regression dependence model:

$$\left(Y \mid X = x\right) = \alpha x + x^{\beta} Z, \text{ for } x > u$$

  - slope parameter α for Y given large X,
  - "spread" parameter $\beta \in (-\infty, 1]$ controls stochasticity of relationship between Y and large X.

Conditional extremes

- Heteroskedastic regression dependence model:

$$\left(\boldsymbol{Y}_{-i} \mid Y_i = y_i\right) = \boldsymbol{\alpha}_{|i} y_i + y_i^{\boldsymbol{\beta}_{|i}} \boldsymbol{Z}_{|i}, \text{ for } y_i > u_{Y_i}$$

  - slope parameter $\alpha_{j|i} \in [-1, 1]$ for Y_j given large Y_i,
  - "spread" parameter $\beta_{j|i} \in (-\infty, 1]$ controls stochasticity of relationship between Y_j and large Y_i.

J. E. Heffernan and J. A. Tawn, *"A Conditional Approach for Multivariate Extreme Values (with Discussion)"* (2004)

Conditional extremes

- **Heteroskedastic regression dependence model**:

$$\left(\mathbf{Y}_{-i} \mid Y_i = y_i\right) = \boldsymbol{\alpha}_{|i}\mathbf{y}_i + \mathbf{y}_i^{\boldsymbol{\beta}_{|i}}\mathbf{Z}_{|i}, \text{ for } y_i > u_{Y_i}$$

  - slope parameter $\alpha_{j|i} \in [-1, 1]$ for Y_j given large Y_i,
  - "spread" parameter $\beta_{j|i} \in (-\infty, 1]$ controls stochasticity of relationship between Y_j and large Y_i.

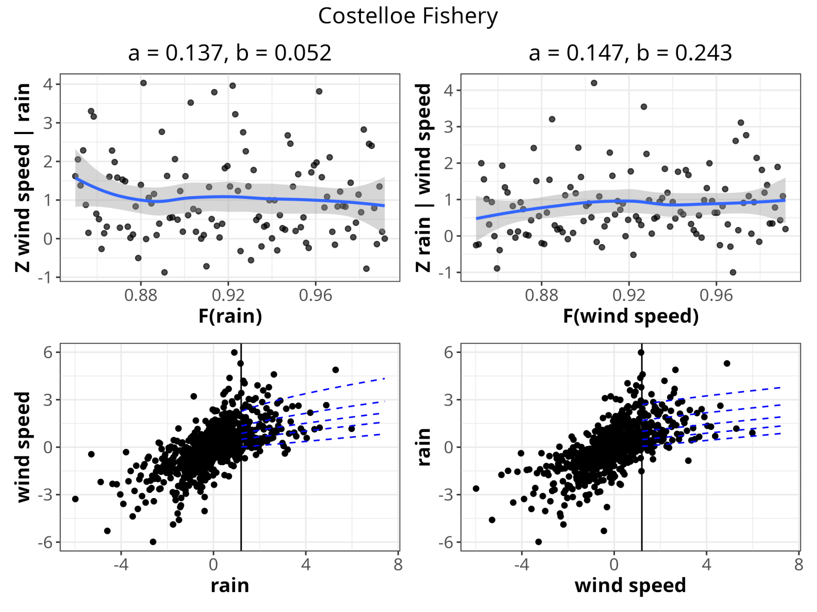
- Key assumptions:
  - Residuals $\mathbf{Z}_{|i} \sim N(\boldsymbol{\mu}_{|i}, \boldsymbol{\Sigma}_{|i})$
  - $\mathbf{Z}_{|i}, Y_i$ conditionally independent for large Y_i

- Special cases:

$$\begin{aligned} \boldsymbol{\alpha}_{|i} = 0, \boldsymbol{\beta}_{|i} = 0 \quad &\Longrightarrow \mathbf{Y}_{-i}, \mathbf{Y}_i \text{ independent,} \\ \boldsymbol{\alpha}_{|i} = -1/1, \boldsymbol{\beta}_{|i} = 0 \quad &\Longrightarrow \text{ perfect positive/negative dependence,} \\ -1 < \boldsymbol{\alpha}_{|i} < 1 \quad &\Longrightarrow \text{ asymptotic independence} \end{aligned}$$

J. E. Heffernan and J. A. Tawn, *"A Conditional Approach for Multivariate Extreme Values (with Discussion)"* (2004)

Conditional extremes



Inference

Inference assumes conditional distribution follows a multivariate Normal (MVN) distribution:

$$\left(\boldsymbol{Y}_{-i} \mid Y_i = y_i\right) \sim N\left(\boldsymbol{\alpha}_{|i} y_i + y_i^{\boldsymbol{\beta}_i} \boldsymbol{\mu}_{|i}, y_i^{\boldsymbol{\beta}_i} \boldsymbol{\Sigma}_{|i}\right), \text{ for } Y_i > u_{Y_i}$$

$\Longrightarrow$ **dependence structures at different sites can be compared using their MVN distributions**

Clustering

skew-geometric Jensen-Shannon divergence

- Kullback-Leibler divergence $KL(X \mid\mid Y)$ measures (**asymmetric**) distance from X to Y
- $KL(X \mid\mid Y)$ has closed form for two MVNs
- $KL(X \mid\mid Y) + KL(Y \mid\mid X)$ symmetric but does not satisfy triangle inequality $\implies$ **not a true metric**.

skew-geometric Jensen-Shannon divergence

- Kullback-Leibler divergence $KL(X \,||\, Y)$ measures (**asymmetric**) distance from X to Y
- $KL(X \,||\, Y)$ has closed form for two MVNs
- $KL(X \,||\, Y) + KL(Y \,||\, X)$ symmetric but does not satisfy triangle inequality $\implies$ **not a true metric**.
- Weighted geometric mean $G_\alpha(X, Y) = X^\alpha Y^{1-\alpha}, \alpha \in [0, 1]$

skew-geometric Jensen-Shannon divergence

- Kullback-Leibler divergence $KL(X \mid\mid Y)$ measures (**asymmetric**) distance from X to Y
- $KL(X \mid\mid Y)$ has closed form for two MVNs
- $KL(X \mid\mid Y) + KL(Y \mid\mid X)$ symmetric but does not satisfy triangle inequality $\implies$ **not a true metric**.
- Weighted geometric mean $G_\alpha(X, Y) = X^\alpha Y^{1-\alpha}, \alpha \in [0, 1]$
- weighted product G_α of two exponential family members is in **same family**.

skew-geometric Jensen-Shannon divergence

- Kullback-Leibler divergence $KL(X \,||\, Y)$ measures (**asymmetric**) distance from X to Y
- $KL(X \,||\, Y)$ has closed form for two MVNs
- $KL(X \,||\, Y) + KL(Y \,||\, X)$ symmetric but does not satisfy triangle inequality $\implies$ **not a true metric**.
- Weighted geometric mean $G_\alpha(X, Y) = X^\alpha Y^{1-\alpha}, \alpha \in [0, 1]$
- weighted product G_α of two exponential family members is in **same family**.
- **Skew-geometric Jensen-Shannon divergence**:

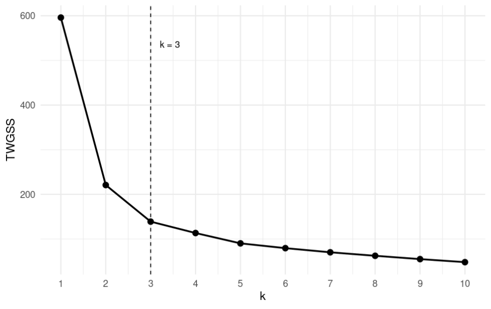
$$JS^{G_\alpha}(X \,||\, Y) = \frac{1}{2}\left\{KL(X \,||\, G_\alpha(X, Y)) + KL(Y \,||\, G_\alpha(X, Y))\right\}$$

Nielsen, *"On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means"* (2019)

Clustering

- Uses **k-medoids** to cluster over JS_{G_α} dissimilarity matrix between sites
- **Elbow plot**: k chosen using **Total Within Group Sum of Squares** between sites x within clusters C_i with medoids m_i:

$$\mathrm{TWGSS}(k) = \sum_{i=1}^{k} \sum_{x \in C_i} JS^{G_\alpha}\left(x \,||\, m_i\right)$$

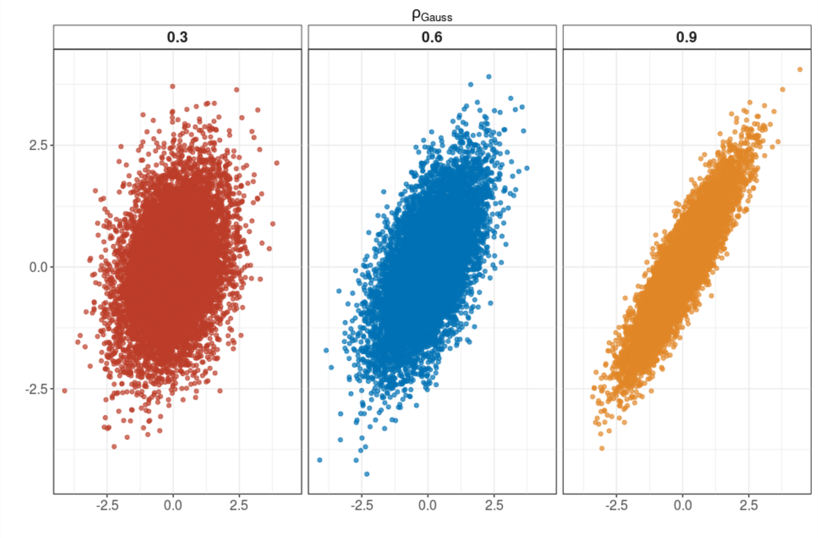


Simulations

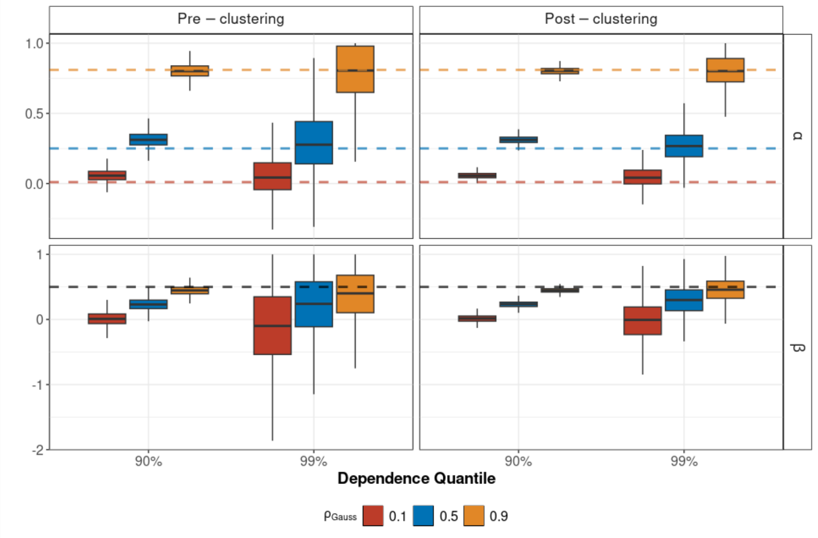
Gaussian copula

- For single "site", generate data from bivariate Gaussian copula with correlation ρ_{Gauss}.
- (asymptotic) CE parameters are $\alpha = \rho^2_{\text{Gauss}}, \beta = 1/2$
- Example: 12 "sites" with $10{,}000$ observations of 2 variables
- 3 clusters of 4 locations defined by respective ρ_{Gauss} values of $0.3, 0.6, 0.9$.

Gaussian copula



Gaussian copula - results



Mixture simulation

- Extend design to mixture of Gaussian and t-copulas

Mixture simulation

- Extend design to mixture of Gaussian and t-copulas
- Idea:
  - Gaussian copula generates observations exhibiting extremal independence,
  - t-copula induces extremal dependence.

Mixture simulation

- Extend design to mixture of Gaussian and t-copulas
- Idea:
  - Gaussian copula generates observations exhibiting extremal independence,
  - t-copula induces extremal dependence.
- Design: 12 sites with $10{,}000$ bivariate observations
- 2 clusters each of 6 sites

Mixture simulation

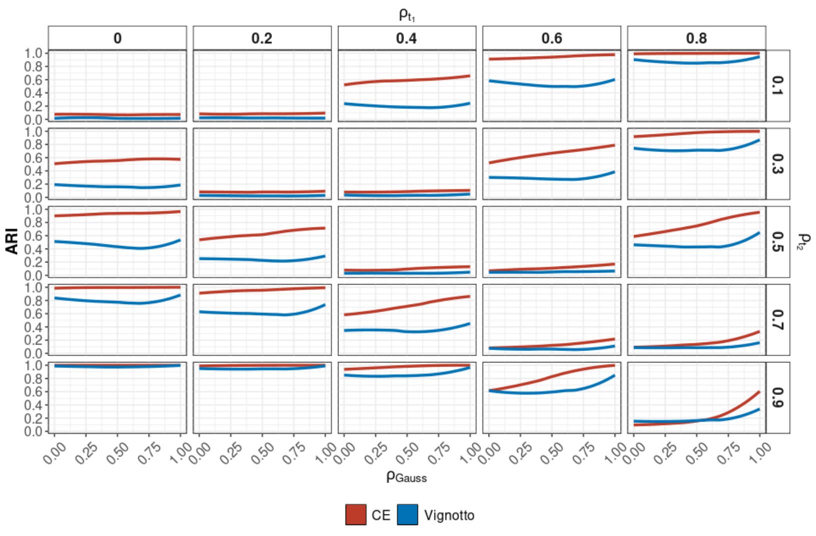
- Extend design to mixture of Gaussian and t-copulas
- Idea:
  - Gaussian copula generates observations exhibiting extremal independence,
  - t-copula induces extremal dependence.
- Design: 12 sites with $10{,}000$ bivariate observations
- 2 clusters each of 6 sites
- Grid search performed over $\rho_{\text{Gauss}}, \rho_{t_1}, \rho_{t_2} \in [0, 1]$.

Mixture simulation

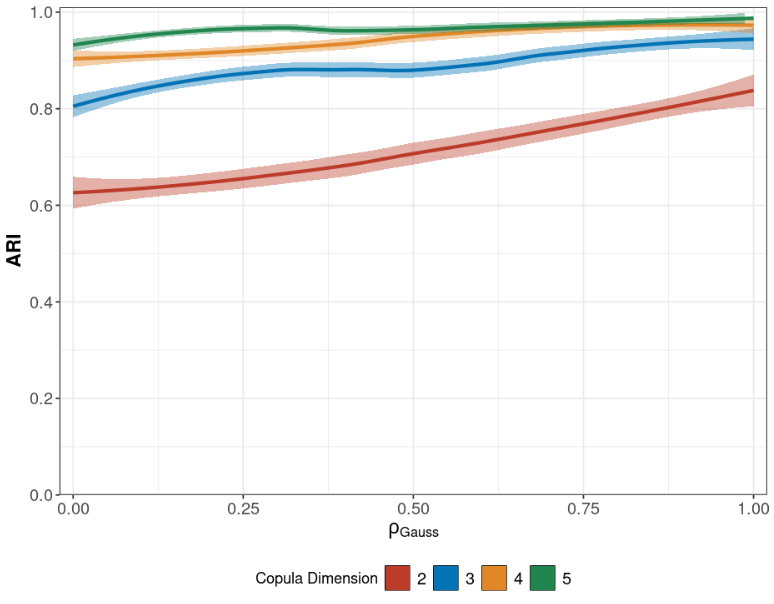
- Extend design to mixture of Gaussian and t-copulas
- Idea:
  - Gaussian copula generates observations exhibiting extremal independence,
  - t-copula induces extremal dependence.
- Design: 12 sites with $10{,}000$ bivariate observations
- 2 clusters each of 6 sites
- Grid search performed over $\rho_{\text{Gauss}}, \rho_{t_1}, \rho_{t_2} \in [0, 1]$.
- Clustering compared to competing Vignotto algorithm using Adjusted Rand Index ARI $\in [0, 1]$

Vignotto, Engelke, and Zscheischler,
"Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland" (2021)

Mixture simulation

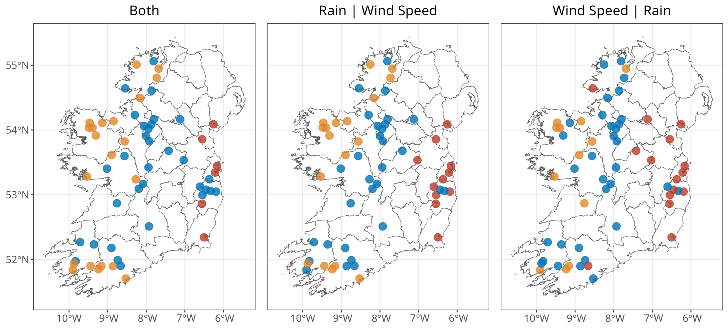


Mixture simulation



Ireland meteorological data

Ireland clustering



Discussion

Discussion

- Conclusion:
  - Principled & effective clustering framework for CE models
  - Simulations & application show clustering can group sites with similar tail dependence, which may aid interpretation

Discussion

- **Conclusion**:
  - Principled & effective clustering framework for CE models
  - Simulations & application show clustering can group sites with similar tail dependence, which may aid interpretation
- **Limitations**:
  - **CE**: Gaussian assumption for $\boldsymbol{Z}_{|i}$
  - **Clustering**: Uncertainty in CE fits is ignored in clustering

Thank you! For slides and supplementary materials, please scan the QR code below:

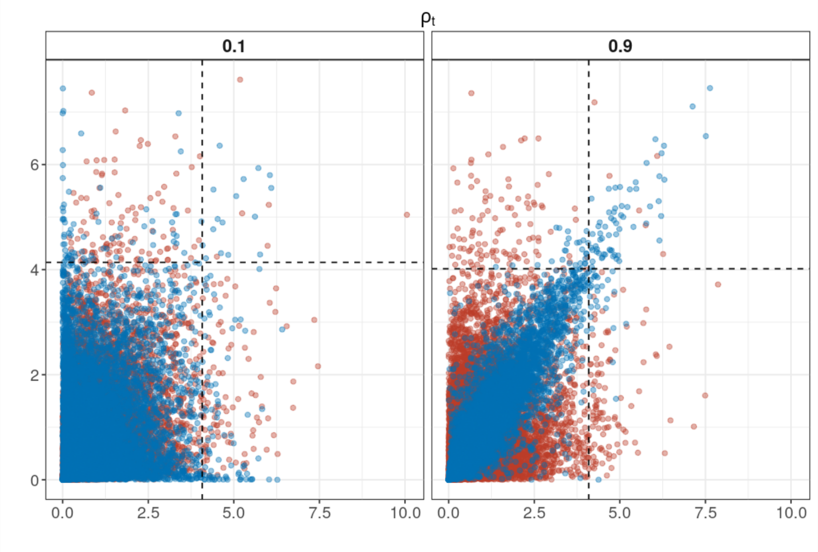
Email: *pot23@bath.ac.uk*

References

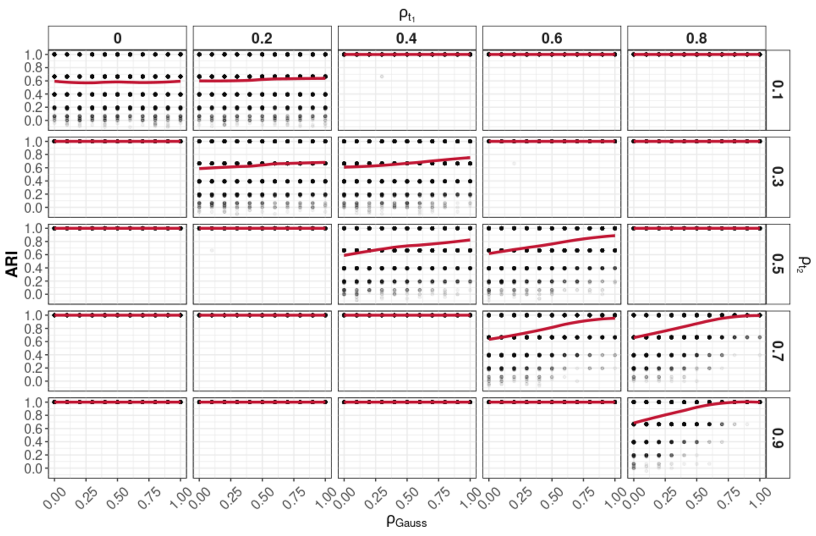
- Coles, Stuart, Janet Heffernan, and Jonathan Tawn (1999). "Dependence measures for extreme value analyses". In: *Extremes (Boston)* 2.4, pp. 339–365.
- Heffernan, Janet E. and Jonathan A. Tawn (July 2004). "A Conditional Approach for Multivariate Extreme Values (with Discussion)". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 66.3, pp. 497–546. ISSN: 1467-9868. DOI: `10.1111/j.1467-9868.2004.02050.x`. URL: `http://dx.doi.org/10.1111/j.1467-9868.2004.02050.x`.
- Nielsen, Frank (May 2019). "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means". In: *Entropy* 21.5, p. 485. ISSN: 1099-4300. DOI: `10.3390/e21050485`. URL: `http://dx.doi.org/10.3390/e21050485`.
- Vignotto, Edoardo, Sebastian Engelke, and Jakob Zscheischler (2021). "Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland". In: *Weather and Climate Extremes* 32, p. 100318.

Appendix

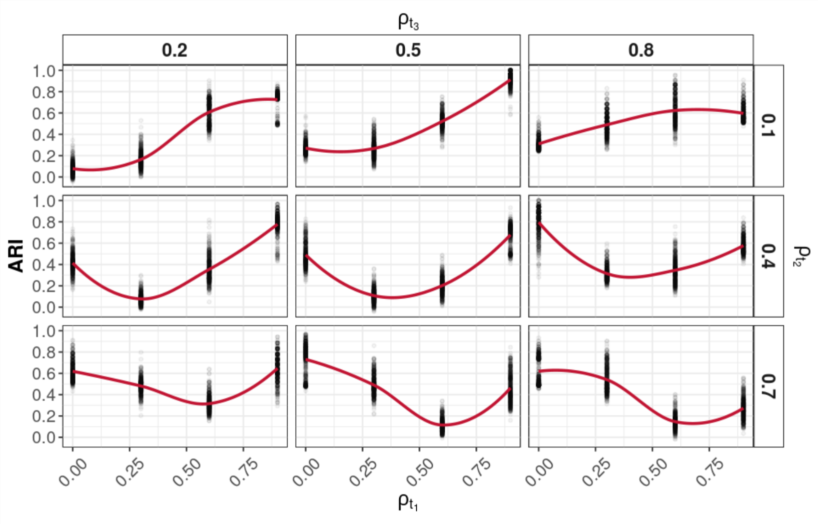
Mixture simulation



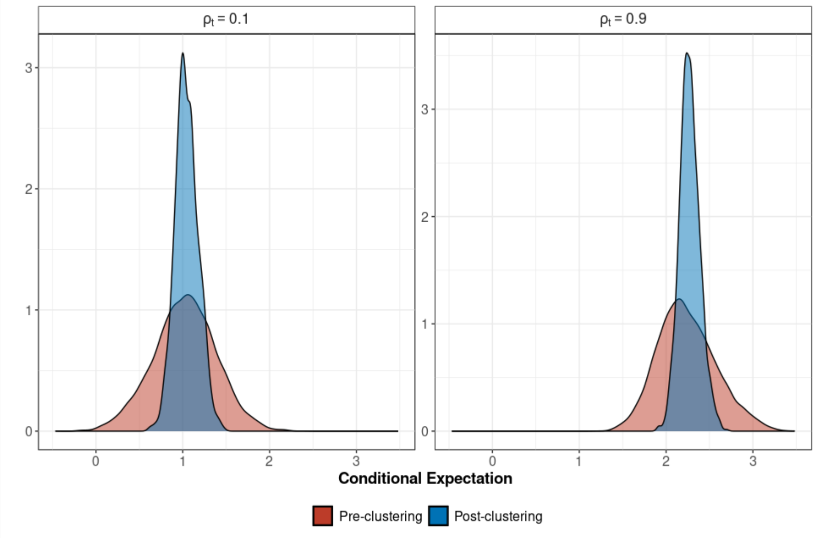
Mixture simulation (three variables)



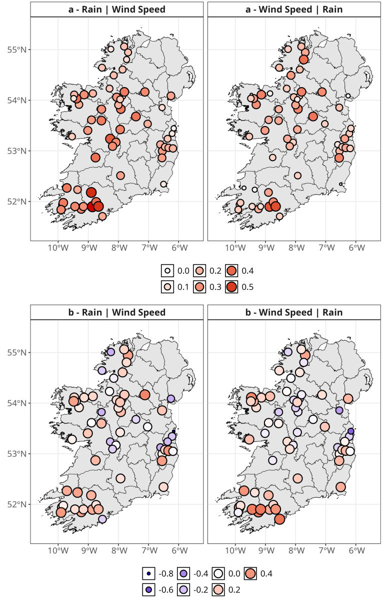
More realistic example



Uncertainty Reduction



α, β pre-clustering



α, β post-clustering