Attributing changes in extreme events

using a causal approach to the tails

Daniela Castro-Camilo and Mengran Li

EXALT workshop 2026

Attributing changes in extreme events

using a causal approach to the tails

Attributing changes in extreme events

using a causal approach to the tails

Attributing changes: Understanding why *things* (extremes) are changing, not just whether they are changing

Attributing changes in extreme events

using a causal approach to the tails

Attributing changes: Understanding why *things* (extremes) are changing, not just whether they are changing

Extreme events: Rare, high-impact events such as heavy prcp, heatwaves, and floods

Attributing changes in extreme events

using a causal approach to the tails

Attributing changes: Understanding why *things* (extremes) are changing, not just whether they are changing

Extreme events: Rare, high-impact events such as heavy prcp, heatwaves, and floods

Causal approach: Estimating how anthropogenic climate change alters extremes*.

Attributing changes in extreme events
using a causal approach to the tails

Attributing changes: Understanding why *things* (extremes) are changing, not just whether they are changing

Extreme events: Rare, high-impact events such as heavy prcp, heatwaves, and floods

Causal approach: Estimating how anthropogenic climate change alters extremes*.

Tails (of distributions): where low-probability extreme events live.

Motivation: the why, the where, and the challenge

The type of problems we're trying to tackle

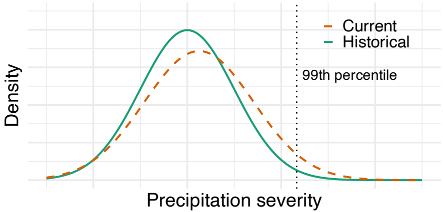
How has anthropogenic climate change altered rare precipitation extremes beyond natural circulation variability?

We address this by comparing very high quantiles of the counterfactual outcome distributions across different climate states (i.e., outcomes under different climate conditions).

How has anthropogenic climate change altered rare precipitation extremes beyond natural circulation variability?

We address this by comparing very high quantiles of the counterfactual outcome distributions across different climate states (i.e., outcomes under different climate conditions). Conceptually, we compare

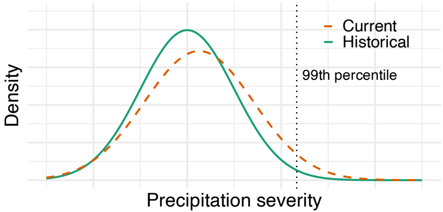
- the historical climate period (1950–1980) and
- the current climate period (1995–2020)



How has anthropogenic climate change altered rare precipitation extremes beyond natural circulation variability?

We address this by comparing very high quantiles of the counterfactual outcome distributions across different climate states (i.e., outcomes under different climate conditions). Conceptually, we compare

- the historical climate period (1950–1980) and
- the current climate period (1995–2020)



Then, after adjusting for atmospheric circulation, we estimate:

Extreme quantile treatment effect

$$\delta(\tau) = Q_{Y_1}(\tau) - Q_{Y_0}(\tau), \qquad \tau \approx 1.$$

Options available in the literature

1. Model the whole distribution

Conditional distribution models, distributional regression, Bayesian hierarchical models.

Risk: tail accuracy depends on getting the whole distribution right.

2. Estimate quantiles directly

Quantile regression, IPW quantile treatment effects, inverse estimating equations.

Risk: pointwise quantile estimation becomes unstable as $\tau \to 1$.

3. Use EVT extrapolation

Hill-type estimators, Pickands-type estimators, tail extrapolation after causal adjustment.

Risk: often heavy-tail-specific and two-stage.

Options available in the literature

1. Model the whole distribution

Conditional distribution models, distributional regression, Bayesian hierarchical models.

Risk: tail accuracy depends on getting the whole distribution right.

2. Estimate quantiles directly

Quantile regression, IPW quantile treatment effects, inverse estimating equations.

Risk: pointwise quantile estimation becomes unstable as $\tau \to 1$.

3. Use EVT extrapolation

Hill-type estimators, Pickands-type estimators, tail extrapolation after causal adjustment.

Risk: often heavy-tail-specific and two-stage.

Our route: directly estimate the causal extreme quantile, but calibrate the tail using EVT.

Goal

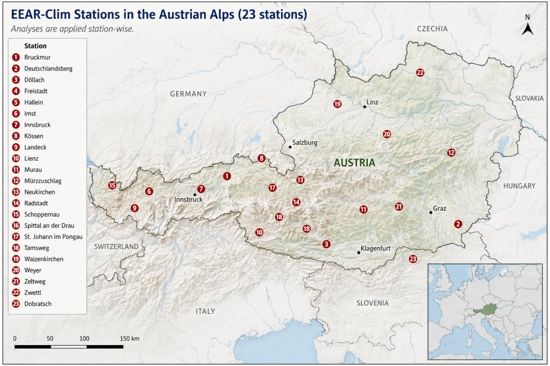
Create a method that combines causal identification and tail extrapolation within a single estimating framework.

- **Outcome**: seasonal 7-day maximum precipitation, analysed separately at each of 23 stations.
- **Data**: EEAR-Clim[a] stations in the Austrian Alps.
- **Treatment contrast**:

$$D = 0 : 1950\text{–}1980, \qquad D = 1 : 1995\text{–}2020.$$

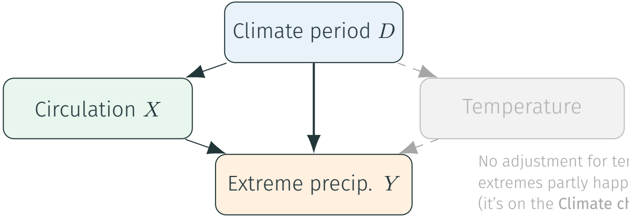
- **Target**: $\tau = 0.99$, approximately a 100-year return-level scale under stationarity.

[a] High-density observational dataset for the Extended European Alpine Region



Question: Conditional on comparable atmospheric circulation, how much has the upper tail of precipitation intensified?

- Atmospheric circulation affects precipitation extremes.
- Circulation patterns also differ between historical and current periods.
- Naive comparisons may mix thermodynamic effects with dynamical variability.



No adjustment for temperature, since effect of anthropogenic warming on rainfall extremes partly happens *through* warmer temperatures (it's on the **Climate change** $\longrightarrow$ **Temp** $\longrightarrow$ **Extreme prcp** causal pathway).

Adjustment set from ERA5: **SLP** (sea-level pressure), **Z500** (geopotential height at 500 hPa), **Wind850** (wind intensity at 850 hPa), **NAO and AO** (large-scale teleconnection indices).

Comparison with standard EEA

What we want to be able to say:

Not just: The probability of exceeding a fixed threshold has changed.

But also: After adjusting for atmospheric circulation, the 99th percentile of 7-day precipitation has increased by about x mm under the current climate.

This is an attribution statement about severity, not only frequency.

Methodology

Methodology based on inverse estimating equations

Estimating equations come from a very general idea in statistics:

Instead of specifying a full likelihood for the data, we estimate parameters by solving equations that should hold "on average" at the true parameter value.

Methodology based on inverse estimating equations

Estimating equations come from a very general idea in statistics:

Instead of specifying a full likelihood for the data, we estimate parameters by solving equations that should hold "on average" at the true parameter value.

Very simply, suppose a parameter θ satisfies

$$E[g(W, \theta)] = 0$$

for some function g

Methodology based on inverse estimating equations

Estimating equations come from a very general idea in statistics:

Instead of specifying a full likelihood for the data, we estimate parameters by solving equations that should hold "on average" at the true parameter value.

Very simply, suppose a parameter θ satisfies

$$E[g(W,\theta)] = 0$$

for some function g

Then, in the data, we replace the expectation by a sample average and solve

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i,\theta) = 0 \quad \leftarrow \text{estimating equation}$$

Methodology based on inverse estimating equations

Estimating equations come from a very general idea in statistics:

Instead of specifying a full likelihood for the data, we estimate parameters by solving equations that should hold "on average" at the true parameter value.

Very simply, suppose a parameter θ satisfies

$$E[g(W, \theta)] = 0$$

for some function g

Then, in the data, we replace the expectation by a sample average and solve

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i, \theta) = 0$$

$\leftarrow$ estimating equation

Why are they useful in causal inference? Because causal problems often involve:

- nuisance params
- weighting
- missing data
- semiparam models

$\rightarrow$ Specifying a full likelihood is difficult.

Methodology based on inverse estimating equations

Estimating equations come from a very general idea in statistics:

Instead of specifying a full likelihood for the data, we estimate parameters by solving equations that should hold "on average" at the true parameter value.

Very simply, suppose a parameter θ satisfies

$$E[g(W, \theta)] = 0$$

for some function g

Then, in the data, we replace the expectation by a sample average and solve

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i, \theta) = 0 \quad \leftarrow \text{estimating equation}$$

Why are they useful in causal inference? Because causal problems often involve:

- nuisance params
- weighting
- missing data
- semiparam models

$\rightarrow$ Specifying a full likelihood is difficult.

Estimating equations allow us to:

- directly target the causal quantity of interest,
- remain robust to partial misspecification,
- avoid modelling the full distribution.

Inverse estimating equations for general QTE estimation

For a potential outcome distribution F_{Y_d},

$$\tau_d(\theta) = \mathbb{P}(Y_d \leq \theta).$$

A quantile is the inverse of this map:

$$Q_{Y_d}(\tau) = \theta \quad \Longleftrightarrow \quad \tau_d(\theta) = \tau.$$

Inverse estimating equations for general QTE estimation

For a potential outcome distribution F_{Y_d},

$$\tau_d(\theta) = \mathbb{P}(Y_d \leq \theta).$$

A quantile is the inverse of this map:

$$Q_{Y_d}(\tau) = \theta \quad \Longleftrightarrow \quad \tau_d(\theta) = \tau.$$

Under causal identification conditions, we construct a signal $s_d(W, \theta, \zeta)$ such that

$$\mathsf{E}\{s_d(W, \theta, \zeta)\} = \mathbb{P}(Y_d \leq \theta).$$

Then estimate θ by solving

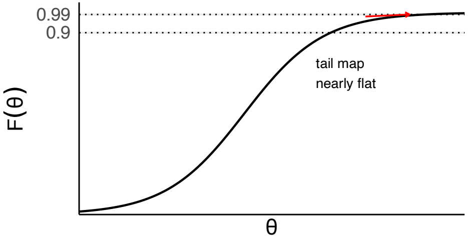
$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\left\{s_d(W_i, \theta, \hat{\zeta}) - \tau\right\}}_{\text{this is } g(W_i, \theta, \hat{\zeta})} = 0.$$

So the signal function is the quantity we average over the data to recover the counterfactual distribution

Why this becomes unstable in the tail

As $\tau \to 1$:

- very few observations inform the root;
- empirical quantiles become sensitive to single observations;
- the density near $Q_Y(\tau)$ may be small;
- standard root-n quantile theory no longer applies.



Effective sample size

When $1-\tau$ is of order $1/n$, the effective sample size in the tail is about one.

Our proposal: Tail-calibrated inverse estimating equations (TIEE)

Replace pointwise inversion by an integrated moment condition

Instead of relying only on information around one extreme quantile, aggregate information across neighbouring quantile levels.

Our proposal: Tail-calibrated inverse estimating equations (TIEE)

Replace pointwise inversion by an integrated moment condition

Instead of relying only on information around one extreme quantile, aggregate information across neighbouring quantile levels.

Define an integrated signal

$$S_d(W,\theta) = \int_0^1 s_{\mathsf{TIEE},d}(W,\theta,\zeta,p)\, dp,$$

chosen so that

$$\mathsf{E}\{S_d(W,\theta)\} = \mathbb{P}(Y_d \le \theta).$$

Our proposal: Tail-calibrated inverse estimating equations (TIEE)

Replace pointwise inversion by an integrated moment condition

Instead of relying only on information around one extreme quantile, aggregate information across neighbouring quantile levels.

Define an integrated signal

$$S_d(W,\theta) = \int_0^1 s_{\text{TIEE},d}(W,\theta,\zeta,p)\,dp,$$

chosen so that

$$\mathsf{E}\{S_d(W,\theta)\} = \mathbb{P}(Y_d \leq \theta).$$

Then estimate the extreme quantile by solving the estimating equation

$$\mathsf{E}\{S_d(W,\theta) - \tau\} = 0.$$

Same causal target, but a more stable route to the tail with an approach that effectively borrows information across nearby quantile levels.

How should we define this causal signal?

- Signal functions are constructed so that their expectation recovers the counterfactual distribution.
- As such, they are not uniquely defined!

How should we define this causal signal?

- Signal functions are constructed so that their expectation recovers the counterfactual distribution.
- As such, they are not uniquely defined!
- A very natural choice is the **naïve indicator signal**

$$s_{\mathsf{TIEE},d}(W,\theta,\zeta,p) = s_{\mathsf{TIEE},d}(\theta,p) = \mathbb{I}\{Q_{Y_d}(p) \le \theta\}.$$

Since $p \in (0,1)$, it "counts" how many quantile levels lie below θ. So integrating over p we get

$$\int_0^1 \mathbb{E}\left[s_{\mathsf{TIEE},d}(W,\theta,\zeta,p)\right]\, dp = \int_0^1 \mathbb{I}\{Q_{Y_d}(p) \le \theta\}\, dp = F_{Y_d}(\theta) = \tau_d(\theta)$$

  - So, on average, this naïve signal recovers the target quantile, **but only in random settings**

How should we define this causal signal?

- Signal functions are constructed so that their expectation recovers the counterfactual distribution.
- As such, they are not uniquely defined!
- A very natural choice is the **naïve indicator signal**

$$s_{\mathsf{TIEE},d}(W,\theta,\zeta,p) = s_{\mathsf{TIEE},d}(\theta,p) = \mathbb{I}\{Q_{Y_d}(p) \leq \theta\}.$$

Since $p \in (0,1)$, it "counts" how many quantile levels lie below θ. So integrating over p we get

$$\int_0^1 \mathbb{E}\left[s_{\mathsf{TIEE},d}(W,\theta,\zeta,p)\right]\, dp = \int_0^1 \mathbb{I}\{Q_{Y_d}(p) \leq \theta\}\, dp = F_{Y_d}(\theta) = \tau_d(\theta)$$

- So, on average, this naïve signal recovers the target quantile, **but only in random settings**
- It doesn't hold in observational studies where treatment assignment depends on covariates.
- In our application, certain circulation patterns may appear more frequently in one climate period than the other, so naïve averages mix circulation effects and warming effects.
- In other words, **atmospheric circulation acts as a confounder**.

How should we define this causal signal?

- To adjust for this confounding, we can instead consider the **propensity score**

$$s_{\mathsf{TIEE},d}(W, \theta, \zeta, p) = \frac{\mathbb{I}(D = d)}{\pi_d(X; \zeta)} \mathbb{I}\{Q_{Y_d}(p) \leq \theta\},$$

where $\pi_d(X; \zeta) = \mathbb{P}(D = d \mid X)$.

Interpretation

The propensity score balances circulation states across climate periods, so the comparison targets a causal contrast under comparable dynamics.

How should we define this causal signal?

- To adjust for this confounding, we can instead consider the **propensity score**

$$s_{\mathsf{TIEE},d}(W,\theta,\zeta,p) = \frac{\mathbb{I}(D=d)}{\pi_d(X;\zeta)}\,\mathbb{I}\{Q_{Y_d}(p) \leq \theta\},$$

where $\pi_d(X;\zeta) = \mathbb{P}(D = d \mid X)$.

Interpretation

The propensity score balances circulation states across climate periods, so the comparison targets a causal contrast under comparable dynamics.

Importantly: our framework is not tied to a specific signal. Other constructions could also be used, including:

- machine learning nuisance estimators,
- continuous-treatment (e.g., global temperature anomaly)

Tail calibration through EVT

- The signal function allows us to reconstruct the counterfactual distribution $F_{Y_d}(\theta)$. So causally, we are now in good shape.

Tail calibration through EVT

- The signal function allows us to reconstruct the counterfactual distribution $F_{Y_d}(\theta)$. So causally, we are now in good shape.
- But a major problem remains: we are interested in $\tau \to 1$ (empirical quantiles become unstable, classical estimating equations break down)

Tail calibration through EVT

- The signal function allows us to reconstruct the counterfactual distribution $F_{Y_d}(\theta)$. So causally, we are now in good shape.
- But a major problem remains: we are interested in $\tau \to 1$ (empirical quantiles become unstable, classical estimating equations break down)
- Even after causal adjustment, we still do not have enough data in the far tail. So we need a principled way to extrapolate beyond the observed extremes.

Tail calibration through EVT

- The signal function allows us to reconstruct the counterfactual distribution $F_{Y_d}(\theta)$. So causally, we are now in good shape.
- But a major problem remains: we are interested in $\tau \to 1$ (empirical quantiles become unstable, classical estimating equations break down)
- Even after causal adjustment, we still do not have enough data in the far tail. So we need a principled way to extrapolate beyond the observed extremes.

For high thresholds u, EVT motivates a GPD approximation for exceedances:

$$Q_{Y_d}(\tau) = u + \frac{\sigma_d}{\xi_d}\left[\left(\frac{1-p_u}{1-\tau}\right)^{\xi_d} - 1\right], \qquad Q_{Y_d}(\tau) > u.$$

Tail calibration through EVT

- The signal function allows us to reconstruct the counterfactual distribution $F_{Y_d}(\theta)$. So causally, we are now in good shape.
- But a major problem remains: we are interested in $\tau \to 1$ (empirical quantiles become unstable, classical estimating equations break down)
- Even after causal adjustment, we still do not have enough data in the far tail. So we need a principled way to extrapolate beyond the observed extremes.

For high thresholds u, EVT motivates a GPD approximation for exceedances:

$$Q_{Y_d}(\tau) = u + \frac{\sigma_d}{\xi_d}\left[\left(\frac{1-p_u}{1-\tau}\right)^{\xi_d} - 1\right], \qquad Q_{Y_d}(\tau) > u.$$

Body ($p \leq p_u$)

Empirical or model-assisted estimation where sufficient data are available.

Tail ($p > p_u$)

EVT-calibrated extrapolation embedded directly in the estimating equation (no two steps).

Steps to implement TIEE

1. Estimate the propensity score $\pi_d(X)$.
2. Choose an intermediate threshold u.
3. Below u: use empirical indicators/data-driven estimation.
4. Above u: fit a GPD tail model for EVT extrapolation.
5. Discretise $(0, 1)$ and approximate the integrated signal over a grid $p_1, \ldots, p_K$.
6. Solve the empirical estimating equation (with some rewritting to express them as a stable convex optimisation problem)

Asymptotics

TIES has the usual desirable asymptotic properties

Identification

The target extreme quantile is the unique root of the population integrated moment.

Consistency

The empirical root converges to the true extreme quantile.

Asymptotic normality

$$\sqrt{n}(\hat{\theta}_{d,\tau} - \theta_{d,\tau}) \Rightarrow N(0, V_{d,\tau}).$$

with extensions for $V_{d,\tau}$ accounting for nuisance parameters (e.g., the params of the propensity score model) estimation.

For eQTE $\delta(\tau) = \theta_{1,\tau} - \theta_{0,\tau}$

we obtain joint asymptotic normality for $(\hat{\theta}_{1,\tau}, \hat{\theta}_{0,\tau})$ and hence

$$\sqrt{n}\{\hat{\delta}(\tau) - \delta(\tau)\} \Rightarrow N(0, \sigma_\delta^2).$$

Main takes from simulation studies

Main takes from simulation studies

1. **Compare against existing methods**, under different tail regimes (heavy- versus light-tailed dist.)

   → TIEE achieves lower bias and MSE[1] and most reliable UQ. Close competitor performs well under heavy-tail. Other estimators don't do very well

2. **Assess robustness** under propensity score misspecification.

   → small bias, comparable MSE against true model

3. **Sensitivity to p_u** (the intermediate quantile level associated with the GPD threshold)

   → MSE varies smoothly without evidence of strong threshold sensitivity

4. **Sensitivity to K**, the grid size used for numerical integration

   → MSE stable across different grid res. and tail regimes

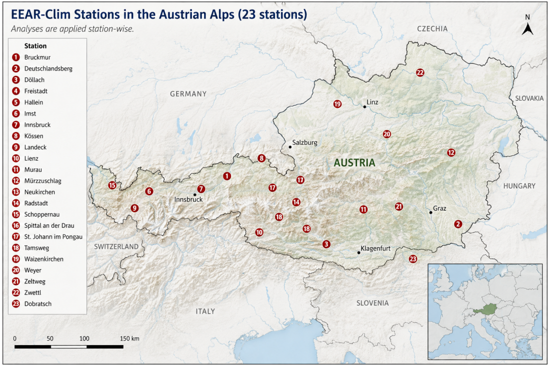
[1] Recall that In EVT, our focus is inferential rather than predictive: we are asking "Can we accurately estimate causal effects on extreme quantiles?". No need for tail-specific scoring rules

Back to the application

- **Outcome**: seasonal 7-day maximum precipitation, analysed separately at each of 23 stations.
- **Treatment contrast**:

$$D = 0 : 1950\text{–}1980, \qquad D = 1 : 1995\text{–}2020.$$

- **Adjustment set:** SLP, Z500, Wind850, NAO, AO.
- **Target quantile level**: $\tau = 0.99$
- **Alternatives (adjusted for confounders):**
  - **Causal Hill estimator:** EVT-based extrapolation tailored to heavy-tailed causal quantiles.
  - **Zhang–Firpo estimator:** semiparametric, based on local tail information



Question: Conditional on comparable atmospheric circulation, how much has the upper tail of precipitation intensified?

Application results: selected station comparisons

Station	Causal Hill	Zhang–Firpo	TIEE	Main message
Landeck	$0.94\ [-1.45, 3.33]$	$-0.20\ [-2.72, 2.93]$	$\mathbf{2.97}\ [\mathbf{1.52}, \mathbf{4.43}]$	TIEE detects significant increase whereas Z-F suggests a decrease
Bruckmur	$1.95\ [-1.98, 5.89]$	$4.20\ [-5.07, 12.70]$	$\mathbf{3.01}\ [\mathbf{0.92}, \mathbf{5.09}]$	ZF produces very wide intervals
Freistadt	$5.04\ [0.51, 9.57]$	$4.30\ [-1.51, 10.27]$	$\mathbf{5.41}\ [\mathbf{2.84}, \mathbf{7.98}]$	Stronger evidence under TIEE
Schoppernau	$8.42\ [3.04, 13.81]$	$11.00\ [0.28, 21.58]$	$\mathbf{5.41}\ [\mathbf{2.86}, \mathbf{7.96}]$	TIEE substantially more precise

Main conclusions

- After adjusting for atmospheric circulation, most stations show positive upper-tail intensification under the modern climate.
- For instance, we can see that the estimated 99th percentile at Landeck increased by about 3 mm under the current climate relative to the historical climate.
- Causal Hill generally captures the correct direction, but with higher variance.
- TIEE yields narrower intervals and more stable detection of significant intensification.

Reflections and extensions

Summary

1. We target eQTE estimation by combining causal adjustment and EVT calibration within one estimating equation.
2. Simulations show improved stability and coverage across tail regimes.
3. In precipitation data, circulation-adjusted estimates reveal a coherent intensification signal.

Where this could go next for climate science

Spatial and spatio-temporal extremes

Embed TIEE within hierarchical spatial EVT models:

- spatially varying tail effects
- Latent Gaussian process structure (SPDE approach to make it fast)

Climate projections

Extend from historical contrasts to future warming scenarios:

- Use CMIP6/DAMIP ensemble simulations to construct factual and counterfactual climate worlds
- Look at continuous treatment (warming level)

Flexible nuisance estimation

Replace parametric nuisance models with ML tools:

- Rather than manually selecting circulation covariates, let ML learn informative circulation summaries from high-dimensional atmospheric fields
- covariate-dependent tail modelling

SCAN ME

Thank you

Key references

- Cheng and Li (2024). Inverting estimating equations for causal inference on quantiles. *Biometrika*.
- Deuber, Li, Engelke and Maathuis (2024). Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *JASA*.
- Firpo (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*.
- Shepherd (2016). A common framework for approaches to extreme event attribution. *Current Climate Change Reports*.
- Zhang (2018a). Extremal quantile treatment effects. *Annals of Statistics*.
- Zhang (2018b). Supplement to "Extremal quantile treatment effects."

Backup slides

Assumptions for causal identification

In observational settings, a few assumptions are needed to make causal inference possible. Specifically,

- Consistency: the observed outcome corresponds to the climate state actually experienced
  - Intuition: if an event occurs under the modern climate, we observe its modern-climate outcome.
- SUTVA (Stable Unit Treatment Value Assumption): One event or season does not affect the outcome of another, and the treatment is well-defined.
  - Intuition: each observation has a clearly defined climate state, with no interference across observations.
- Unconfoundedness: $(Y_0, Y_1) \perp D \mid X$. After adjusting for atmospheric circulation, the climate period assignment is independent of the potential outcomes.
  - Intuition: once circulation patterns are accounted for, differences in extremes can be attributed to warming rather than to different weather regimes.
- Overlap: $0 < \mathbb{P}(D = 1 \mid X) < 1$. For every circulation state, both historical and modern climate conditions are possible.
  - Intuition: we must be able to compare similar atmospheric situations across climate periods.

Caveats:

- Consistency and SUTVA are usually conceptual/design assumptions; they cannot be directly tested from the data.
- Unconfoundedness is fundamentally untestable because it concerns unobserved counterfactuals. We can only make it more plausible by adjusting for scientifically relevant variables (e.g. atmospheric circulation).
- Overlap is the one assumption we can partially assess empirically, for example by checking propensity score distributions.

More on simulations – Setup

Goals:

1. Compare the inverse probability weighted TIEE estimator (TIEE-IPW) with existing methods
2. Examine performance across different tail regimes (heavy- versus light-tailed distributions)
3. Assess robustness of TIEE-IPW under propensity score misspecification.
4. Sensitivity to the intermediate quantile level p_u associated with the GPD threshold
5. Sensitivity to K, the grid size used for numerical integration

Setup

1. We generate n i.i.d. obs. $W(i) = (Y(i), D(i), \boldsymbol{X}(i))$, where $\boldsymbol{X}(i) = (1, X(i))^\top$ and $X(i) \sim \text{Unif}(-1, 1)$.
2. The propensity score for the binary treatment $D = 1$ is

$$\pi(\boldsymbol{x}) = \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.5x^2 + 0.25.$$

3. Treatment assignments are generated as $D(i) \mid \boldsymbol{X}(i) \sim \text{Ber}(\pi(\boldsymbol{X}(i)))$
4. The potential outcomes Y_1 and Y_0 follow one of two tail scenarios: heavy and light
5. Target tail prob: $(1 - \tau_n) \in \{5/n,\ 1/n,\ 5/(n \log n)\}$ (intermediate, moderately extreme, very extreme)
6. Following Deuber et al., 2024, the intermediate quantile level is set to $(1 - p_u) = k/n$ with $k = n^{0.65}$.
7. $n = 1000, 5000$
8. 1000 Monte Carlo replications

Heavy-tail scenario

Following Deuber et al., 2024:

$$M_1^{(\mathrm{H})} : \begin{cases} Y_1 = 5S(1+X), \\ Y_0 = S(1+X), \end{cases} \quad M_2^{(\mathrm{H})} : \begin{cases} Y_1 = C_2 \exp(X), \\ Y_0 = C_3 \exp(X), \end{cases} \quad M_3^{(\mathrm{H})} : \begin{cases} Y_1 = P_{1.75+X,2}, \\ Y_0 = P_{1.75+X,1}, \end{cases}$$

$S \sim t(3)$, $C_s \sim$ Fréchet$(0, 1, s)$, and $P_{a,b} \sim$ Pareto(a, b)

Light-tail scenario

$$M_1^{(\mathrm{L})} : \begin{cases} Y(1) = 5S(1+X), \\ Y(0) = S(1+X), \end{cases} \quad M_2^{(\mathrm{L})} : \begin{cases} Y(1) = C_1 \exp(X), \\ Y(0) = C_2 \exp(X), \end{cases} \quad M_3^{(\mathrm{L})} : \begin{cases} Y(1) = W_{2+X,\,2}, \\ Y(0) = W_{3+2X,\,1}. \end{cases}$$

$S \sim t(3)$, $C_1 \sim$ Exp(1), $C_2 \sim$ Exp(2), and $W_{a,b} \sim$ Weibull(a, b)

More on simulations – Results for $n = 1000$

Heavy-tailed scenarios							
		Scenario $M_1^{(H)}$		Scenario $M_2^{(H)}$		Scenario $M_3^{(H)}$	
$1-\tau_n$	Method	Bias	MSE	Bias	MSE	Bias	MSE
$5/(n\log n)$	Causal Hill	45.982	4828.099	6.370	1355.930	26.759	3652.382
	Pickands	-52.322	5694.769	42.410	56223.057	63.741	86214.119
	Zhang-Firpo	-3.243	2840.382	0.847	5395.063	-9.793	6191.232
	TIEE-IPW	-6.792	**1513.337**	-16.460	**581.743**	-5.879	**544.757**
$1/n$	Causal Hill	28.305	2329.556	4.388	844.775	15.660	1872.560
	Pickands	-36.222	6194.962	59.070	110234.161	58.364	101126.968
	Zhang-Firpo	4.117	2846.812	8.653	5469.215	-2.782	6103.073
	TIEE-IPW	-4.997	**1025.129**	-10.776	**390.789**	-4.072	**390.293**
$5/n$	Causal Hill	4.496	135.273	0.539	55.963	2.142	79.982
	Pickands	-19.168	899.346	6.520	1714.555	4.988	1268.937
	Zhang-Firpo	1.194	174.980	1.183	131.197	3.330	1128.588
	TIEE-IPW	-1.193	**101.772**	-2.066	**52.606**	-1.680	**68.289**

- TIEE-IPW achieves lower bias and variance across all heavy-tail regimes.
- Causal Hill performs well in intermediate regimes, but TIEE-IPW remains more stable in the far tail.
- Pickands exhibits large finite-sample variability, while Zhang–Firpo deteriorates when treatment groups exhibit increasingly different extreme-tail behaviour.

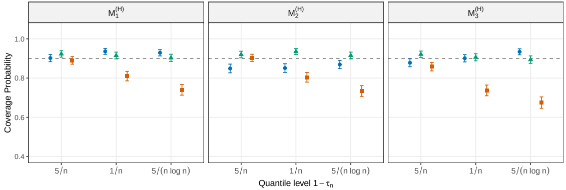
More on simulations – Comparison with existing methods, heavy/light regimes, $n = 1000$

Light-tailed scenarios							
		Scenario $M_1^{(\mathrm{L})}$		Scenario $M_2^{(\mathrm{L})}$		Scenario $M_3^{(\mathrm{L})}$	
$1-\tau_n$	Method	Bias	MSE	Bias	MSE	Bias	MSE
$5/(n\log n)$	Causal Hill	13.488	244.970	4.372	49.431	0.535	0.601
	Pickands	-15.687	482.247	5.818	3791.332	-1.827	15.487
	Zhang-Firpo	-0.337	14.195	0.475	12.403	-0.347	0.307
	TIEE-IPW	-1.286	**10.593**	-0.153	**9.121**	-0.097	**0.252**
$1/n$	Causal Hill	10.916	163.065	3.459	32.668	0.441	0.447
	Pickands	-15.195	408.884	3.743	2018.179	-1.788	12.386
	Zhang-Firpo	0.406	14.245	0.847	12.895	-0.266	0.257
	TIEE-IPW	-1.144	**8.332**	-0.109	**6.573**	-0.089	**0.196**
$5/n$	Causal Hill	2.563	12.662	0.734	3.189	0.105	0.083
	Pickands	-11.699	183.880	-0.564	106.240	-1.438	4.721
	Zhang-Firpo	-0.090	4.283	0.058	2.774	-0.015	0.089
	TIEE-IPW	-0.488	**2.484**	0.173	**1.334**	-0.063	**0.050**

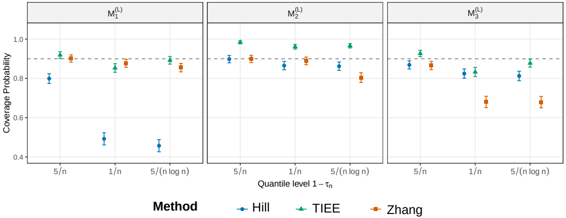
- All estimators perform better in light-tailed settings, with substantially lower MSE.
- TIEE-IPW still performs best, although its advantage is less pronounced.
- Pickands again performs worst, particularly in the very extreme tail.

Coverage rates of 90% confidence intervals for the extreme quantile treatment effect $\delta(\tau_n)$

Heavy-tailed scenarios



Light-tailed scenarios



Results for Pickands estimator are omitted due to highly unstable interval estimates.

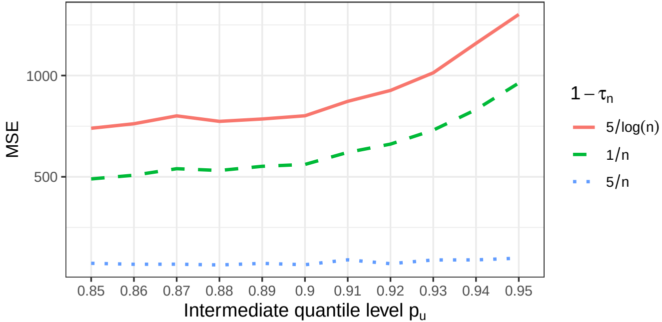
More on simulations – Results to propensity score misspecification

Bias and MSE of the TIEE estimator of the EQTE $\delta(\tau_n)$ under different propensity score misspecifications

$1 - \tau_n$	True Value	Propensity Score Model	Bias	MSE
$5/(n \log n)$	45.39	True Model (Quadratic)	-5.438	374.19
		Polynomial Basis	-5.601	397.00
		Misspecified Form (Linear)	-3.903	386.64
		Misspecified Link (Logit)	-3.269	551.88
		Spurious Covariates	-3.472	581.16
$1/n$	38.38	True Model (Quadratic)	-3.837	275.01
		Polynomial Basis	-3.838	283.39
		Misspecified Form (Linear)	-2.516	291.51
		Misspecified Link (Logit)	-2.199	361.22
		Spurious Covariates	-2.103	370.79
$5/n$	16.53	True Model (Quadratic)	-1.338	29.53
		Polynomial Basis	-1.354	29.33
		Misspecified Form (Linear)	-1.009	31.11
		Misspecified Link (Logit)	-1.287	29.74
		Spurious Covariates	-1.285	29.89

More on simulations – Sensitivity to threshold level p_u

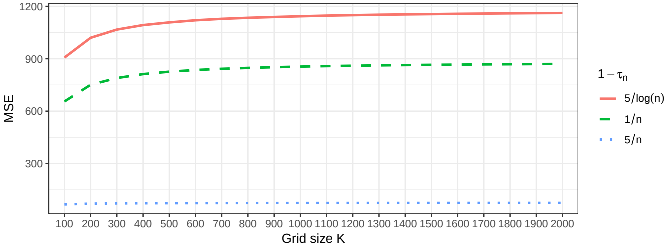
MSE of the TIEE-IPW estimator of the EQTE $\delta(\tau_n)$ as a function of the intermediate quantile level p_u



- As expected, MSE increases as we move further into the tail
- Performance varies smoothly without evidence of strong threshold sensitivity

More on simulations – Sensitivity to grid size K

MSE of the TIEE-IPW estimator of the EQTE $\delta(\tau_n)$ as a function of the grid size K



- MSE remains stable across different grid resolutions and tail regimes.
- Relatively coarse grids already provide sufficient numerical accuracy.
- Results suggest that the integrated estimating equation is numerically stable and robust to discretisation choices.

TIEE achieves practical robustness, not double robustness

DR means consistency is retained if either the confounding model or the tail model is correctly specified.
TIEE is not formally doubly robust, but empirically shows good robustness to misspecification.