Supplementary Information

Supplementary Methods

Supplementary Method 1: Automated threshold selection

To ensure reproducibility and to avoid subjectivity in defining extreme event thresholds, we provide here a detailed description of the automated threshold selection procedure applied in this study. A robust and unbiased definition of thresholds is essential for consistent identification across indicators, and this supplementary methods outlines the procedure used to derive statistically indistinguishable thresholds. The automated selection process consists of the following steps, with detailed parameter settings given in Supplementary Tab. 4:

1. **Tail Transformation:** System stress indicators I are transformed so that system-critical events reside in the right (upper) tail of the distribution. For example, wind capacity factors (where low values indicate stress) are inverted so that extreme events correspond to high values:

$$\boxed{I' = \begin{cases} -I, & \text{if required (e.g., wind CF)} \\ I, & \text{otherwise} \end{cases}} \tag{1}$$

2. **Knee-Point Analysis:** The empirical distribution of each indicator after tail transformation I' is analyzed to detect the point of highest curvature ("knee"), indicating the onset of tail behavior. Thresholds are expressed in the quantile space to make results comparable across indicators and time periods. Following common practice in extreme-value analysis for energy systems [16], we set a conservative lower bound for the candidate thresholds as the maximum of (i) the knee-point quantile rounded up to the next integer percentile and (ii) a minimum baseline percentile $q_{\min}$. The lower quantile threshold τ_{lower} is then defined as:

$$\boxed{\tau_{\text{lower}} = \max\left\{\frac{\lceil 100 \cdot q_{\text{knee}} \rceil}{100},\ q_{\min}\right\}} \tag{2}$$

   Here, $q_{\text{knee}} \in [0, 1]$ denotes the quantile level corresponding to the knee-point of the empirical distribution of I', and $q_{\min}$ is the minimal admissible percentile. In this study we set $q_{\min}$ to the 95^{th} percentile, following common practice in extreme-value analysis [16, 30].
3. **Candidate Threshold Grid:** Starting from the minimum quantile threshold τ_{lower}, candidate quantile thresholds τ are defined on a fine quantile grid $\mathcal{T}_{\text{candidate}}$ with step size $\Delta_\tau = 0.001$:

$$\boxed{\mathcal{T}_{\text{candidate}} = \left\{\tau \;\middle|\; \tau_{\min} \le \tau \le 1.0 - \Delta_\tau,\ \tau = \tau_{\min} + k \cdot \Delta_\tau,\ k \in \mathbb{N}\right\}} \tag{3}$$

4. **Z-Normalization:** Depending on the threshold type thr, Z-normalization is applied element-wise over the time series to ensure comparability across scales and years y:

$$\boxed{Z_t = \begin{cases} \dfrac{I'_t - \mu}{\sigma}, & \text{for } thr \notin \{\text{yearly}\} \\ \dfrac{I'_{t,y} - \mu_y}{\sigma_y}, & \text{for } thr \in \{\text{yearly}\} \end{cases}} \tag{4}$$

   Here, I'_t and $I'_{t,y}$ denote the indicator at timestep t (global) and at timestep t in year y, respectively. μ and σ are the global mean and standard deviation, while μ_y and σ_y are calculated individually per year.

5. **GPD Fitting with Bootstrapping:** To approximate the tail behavior above each threshold, we repeatedly resample the exceedances and fit a theoretical tail model to capture its shape and uncertainty. For each candidate threshold τ, exceedances X quantify how much the values of the standardized indicator Z exceed the corresponding threshold value z_τ and are calculated as:

$$\boxed{\begin{aligned} X &= \{Z - z_\tau \mid Z > z_\tau\}, \\ \text{where } z_\tau &= \text{Quantile}_\tau(Z) \end{aligned}} \tag{5}$$

   Here, $\text{Quantile}_\tau(Z)$ denotes the τ-quantile of Z, i.e., the value below which a fraction τ of the data lies. These exceedances are repeatedly resampled with replacement using k bootstrap replicates. A Generalized Pareto Distribution (GPD) is then independently fitted to each bootstrap sample to model tail behavior and quantify parameter uncertainty.

6. **Distance Metric:** To measure how the fitted theoretical tail model matches the empirical tail, we calculate the average absolute difference between their quantiles across several probability levels. For each bootstrap sample k, the mean absolute distance $\overline{d}_k(\tau)$ between empirical (EMP) and theoretical quantiles is calculated over m evenly spaced probability levels p_i:

$$\boxed{\overline{d}_k(\tau) = \frac{1}{m} \sum_{i=1}^{m} |Q_{\text{EMP}}(p_i) - Q_{\text{GPD}}(p_i)|} \tag{6}$$

   Here, $p_i = \frac{i}{m+1}$ for $i = 1, \ldots, m$ define the probability levels used for matching. $Q_{\text{emp}}(p_i)$ denotes the empirical quantile of the resampled exceedances at level p_i, and $Q_{\text{GPD}}(p_i)$ denotes the theoretical quantile from the GPD fitted to the bootstrap sample.

7. **Threshold Filtering:** For each candidate threshold τ, the mean Anderson-Darling (AD) statistic across bootstrap replicates is used to evaluate the quality of GPD fits. Thresholds are retained only if the average p-value of the AD test $\overline{p}_{\text{AD}}$ exceeds 0.05 and the number of identified extreme events $\mathcal{N}_e(\tau)$ is at least

62 (ensuring one event per year in the design period):

$$\boxed{\mathcal{T}_{\text{filtered}} = \left\{\tau \in \mathcal{T}_{\text{candidate}} : \ \overline{p}_{\text{AD}}(\tau) > 0.05 \ \wedge \ \mathcal{N}_e(\tau) \geq 62\right\}} \tag{7}$$

Here, $N_e(\tau)$ denotes the number of distinct extreme events identified for the threshold τ, as defined by the sequent peak algorithm. The AD test is chosen over the Kolmogorov–Smirnov (KS) test because it gives greater weight to the tails of the distribution, which is essential for extreme value modeling focused on rare, high-impact events.

8. **Optimal Threshold Selection:** The optimal quantile threshold τ_{opt} is defined as the one minimizing the mean distance $\langle \overline{d}_k(\tau) \rangle_k$ across bootstrap replicates among the candidate thresholds that pass filtering:

$$\boxed{\tau_{opt} = \arg \min_{\tau \in \mathcal{T}_{\text{filtered}}} \langle \overline{d}_k(\tau) \rangle_k} \tag{8}$$

9. **Defining Threshold Ranges:** To account for sampling uncertainty and avoid overfitting, we define an indistinguishable quantile threshold range $\mathcal{T}^*$ around the optimal threshold τ_{opt}. This range includes all thresholds forming a contiguous block around τ_{opt} whose average distances $\langle \overline{d}_k(\tau) \rangle_k$ lie within the bootstrap-derived one-sigma confidence interval around τ_{opt}:

$$\boxed{\mathcal{T}^* = \left\{\tau \in \mathcal{T}_{\text{filtered}} : \ \langle \overline{d}_k(\tau) \rangle_k \in CI_{1\sigma}\left(\langle \overline{d}_k(\tau_{opt}) \rangle_k\right)\right\}} \tag{9}$$

Here, $CI_{1\sigma}$ denotes the central 68 % bootstrap confidence interval, defined as the interval between the 16th and 84th percentiles of the bootstrap distribution of $\langle \overline{d}_k(\tau_{opt}) \rangle_k$. This one-sided yet compact interval conservatively captures the threshold variability around the optimum τ_{opt}, avoiding overly broad threshold ranges seen with symmetric intervals [32].

Supplementary Method 2: Extreme event identification method

To consistently capture both the duration and severity of stress periods, we extend the sequent peak algorithm into a severity-aware extreme event identification method. This supplementary method details the procedure, which integrates results across multiple thresholds and applies a probabilistic weighting to emphasize rare and operationally relevant extremes. The resulting method consists of the following steps:

1. **Event Mask Construction:** For each threshold τ^*, a binary event mask $\mathbf{1}_{t,\tau^*}$ is created:

$$\boxed{\mathbf{1}_{t,\tau^*} = \begin{cases} 1, & \text{if } CD^{\text{SPA}}_{t,\tau^*} > 0, \\ 0, & \text{otherwise.} \end{cases}} \tag{10}$$

This mask marks all timesteps t that are part of an extreme event as identified by the SPA-algorithm for the threshold τ^*. These binary masks serve as the foundation for integrating results across multiple thresholds.

2. **Extreme Event Identification:** For each threshold τ^*, the set of identified extreme events $\mathcal{E}_{\tau^*}$ is defined as the collection of all maximal contiguous intervals of timesteps t for which the event mask $\mathbf{1}_{t,\tau^*}$ equals one:

$$\boxed{\mathcal{E}_\tau = \{e^1_{\tau^*}, e^2_{\tau^*}, \ldots, e_{\tau^* u^{N_{\tau^*}}}\}, \quad \text{with} \quad e^i_{\tau^*} = [t_s^{(i)}, t_e^{(i)}] \subseteq \mathbb{T}} \tag{11}$$

Each event $e^i_{\tau^*}$ corresponds to a contiguous time interval where $\mathbf{1}_{t,\tau^*} = 1$ for all $t \in [t_s^{(i)}, t_e^{(i)}]$, and $\mathbf{1}_{t,\tau^*} = 0$ for $t = t_s^{(i)} - 1$ and $t = t_e^{(i)} + 1$, ensuring that the events are maximal.

3. **Exceedance Probability Weighting:** For each threshold τ^*, compute its overall exceedance probability EP_{τ^*} (the fraction of hours classified as extreme) and derive weights w_{τ^*} that penalize thresholds identifying too many hours as extreme:

$$\boxed{w_\tau = -\log(EP_{\tau^*})} \tag{12}$$

This weighting scheme emphasizes thresholds that isolate rarer, more selective extreme conditions, with the logarithmic form ensuring smoother scaling across exceedance probabilities.

Supplementary Method 3: Extreme event quantification metrics

To evaluate extreme events in a consistent and comparable manner, we define a set of severity-, duration-, and impact-based metrics that integrate information across multiple thresholds. This supplementary note introduces the quantification framework and its logarithmic weighting scheme, which together provide a comprehensive characterization of extreme events.

Extreme events are quantified using a logarithmic-weighted average of threshold-specific metrics $M_{t,\tau}$ across the threshold range $\mathcal{T}^*$:

$$\boxed{\langle M_t \rangle_{w_{\tau^*}} = \frac{\sum_{\tau^* \in \mathcal{T}^*} w_{\tau^*} M_{t,\tau^*}}{\sum_{\tau^* \in \mathcal{T}^*} w_{\tau^*}}} \tag{13}$$

Here, M_{t,τ^*} denotes the per-threshold value of a chosen metric (e.g., severity, duration, Consumer Cost), and w_{τ^*} are the exceedance-probability weights. From this general formulation, the following extreme event quantification metrics are defined, with subscript e for event-level values, t for time series, and no subscript for metrics aggregated over all events:

- **Integrated Probability:** The probability of how consistently each timestep t is classified as extreme across thresholds:

$$\boxed{P_t = \langle \mathbf{1}_{t,\tau^*} \rangle_{w_\tau^*}} \tag{14}$$

  Here, $P_t \in [0, 1]$ represents a probability-like measure indicating the weighted share of thresholds that classify timestep t as extreme, thereby reflecting the degree of cross-threshold agreement. While P_t is primarily used for visualization and qualitative interpretation, it may also serve as a filtering criterion to identify time intervals that are consistently recognized as extreme across multiple thresholds. In this work, we apply a zero-threshold filter ($P_t > 0$) to visualize all timesteps that are classified as extreme by at least one threshold.
- **Frequency:** The total number of extreme events, averaged across thresholds and floored:

$$\boxed{F = \left\lfloor \langle N_e(\tau^*) \rangle_{w_{\tau^*}} \right\rfloor} \tag{15}$$

  Here, $N_e(\tau^*)$ is the number of events detected at threshold τ^*.
- **Severity:** The maximum weighted SPA-based deficit within an event:

$$\boxed{S_e = \max_{t \in e} \langle CD^{\mathrm{SPA}}_{t,\tau^*} \rangle_{w_{\tau^*}}} \tag{16}$$

- **Duration:** The weighted sum of extreme event timesteps, representing the event's total length:

$$\boxed{D_e = \sum_{t \in e} \langle \mathbf{1}_{t,\tau^*} \rangle_{w_{\tau^*}} \cdot \Delta t} \tag{17}$$

  Here, $\mathbf{1}_{t,\tau^*}$ is the binary event mask, and Δt is the timestep length.
- **Buildup:** The duration of the extreme event build up, from start to maximum severity:

$$\boxed{B_e = \sum_{t \le t_{\mathrm{peak}} \in e} \langle \mathbf{1}_{t,\tau^*} \rangle_{w_{\tau^*}} \cdot \Delta t} \tag{18}$$

  Here, $t_{\mathrm{peak}} \in e$ is the timestep within the event e with maximum severity S_e.
- **Recovery:** The duration of the recovery phase of the extreme event, from maximum severity to the end of the event:

$$\boxed{R_e = \sum_{t > t_{\mathrm{peak}} \in e} \langle \mathbf{1}_{t,\tau^*} \rangle_{w_{\tau^*}} \cdot \Delta t} \tag{19}$$

- **Event Consumer Cost:** The accumulated Consumer Cost during an extreme event:

$$\boxed{C_e = \sum_{t \in e} \langle \mathrm{CC}_{t,\tau^*} \rangle_{w_{\tau^*}}} \tag{20}$$

- **Event Unmet Energy Demand:** The accumulated *Unmet Energy Demand* covered by an extreme event:

$$\boxed{U_e = \sum_{t \in e} \langle \sum_i UED_{i,t,\tau^*} \rangle_{w_{\tau^*}}} \tag{21}$$

- **Coverage Ratios:** The share of a given metric's total value that is covered by an extreme event:

$$\boxed{Cov_e = \frac{\sum_{t \in e} \langle \mathrm{M}_{t,\tau^*} \rangle_{w_{\tau^*}}}{\sum_t \mathrm{M}_t}} \tag{22}$$

All metrics are computed per event, aggregated annually, and across the full dataset, providing a comprehensive multi-threshold characterization of extreme events.

Supplementary Method 4: Indicator accuracy quantification methods

To evaluate how reliably different indicators reproduce benchmark extreme events, this note defines a set of complementary accuracy metrics. These measures quantify precision, recall, temporal overlap, and alignment quality, providing a robust basis for comparing indicator performance in event-based identification. The quantification metrics are defined as follows:

- **Event Precision:** Fraction of predicted events that are correctly identified by at least one matching benchmark event.

$$\boxed{\mathrm{Precision} = \left\langle \frac{|\mathcal{TP}_{I,\tau^*}|}{|\mathcal{E}_{I,\tau^*}|} \right\rangle_{w_{\tau^*}}} \tag{23}$$

  The precision indicates how many predicted events match actual benchmark events. A high precision means that the indicator produces few false positives.
- **Event Recall:** Fraction of benchmark events that are correctly identified by at least one matching predicted event.

$$\boxed{\mathrm{Recall} = \left\langle \frac{|\mathcal{TP}_{B,\tau^*}|}{|\mathcal{E}_{B,\tau^*}|} \right\rangle_{w_{\tau^*}}} \tag{24}$$

  The recall indicates how many actual benchmark events were detected. A high recall means few true events are missed by the indicator.
- **Event F1-Score:** Harmonic mean of precision and recall.

$$\boxed{\mathrm{F1} = \left\langle \frac{2 \cdot \mathrm{Precision}_{\tau^*} \cdot \mathrm{Recall}_{\tau^*}}{\mathrm{Precision}_{\tau^*} + \mathrm{Recall}_{\tau^*}} \right\rangle_{w_{\tau^*}}} \tag{25}$$

The F1-score balances both precision and recall, providing a single metric that equally penalizes both false positives and false negatives.

- **Benchmark Overlap:** Average relative overlap across all matched benchmark events.

$$\boxed{\text{Overlap} = \left\langle \frac{1}{|\mathcal{M}_\tau|} \sum_{(i,j)\in\mathcal{M}_{\tau^*}} r^B_{ij,\tau^*} \right\rangle_{w_{\tau^*}}} \tag{26}$$

The benchmark overlap indicates how strongly the benchmark events align with the predicted events. A high overlap means that the correctly identified benchmark events are also well covered in time. For example, a value close to one suggests that matched benchmark events are almost fully overlapped.

- **Overlap Count:** Number of event pairs that exhibit any non-zero temporal overlap.

$$\boxed{\text{Count} = \langle |\mathcal{M}_{\tau^*}| \rangle_{w_{\tau^*}}} \tag{27}$$

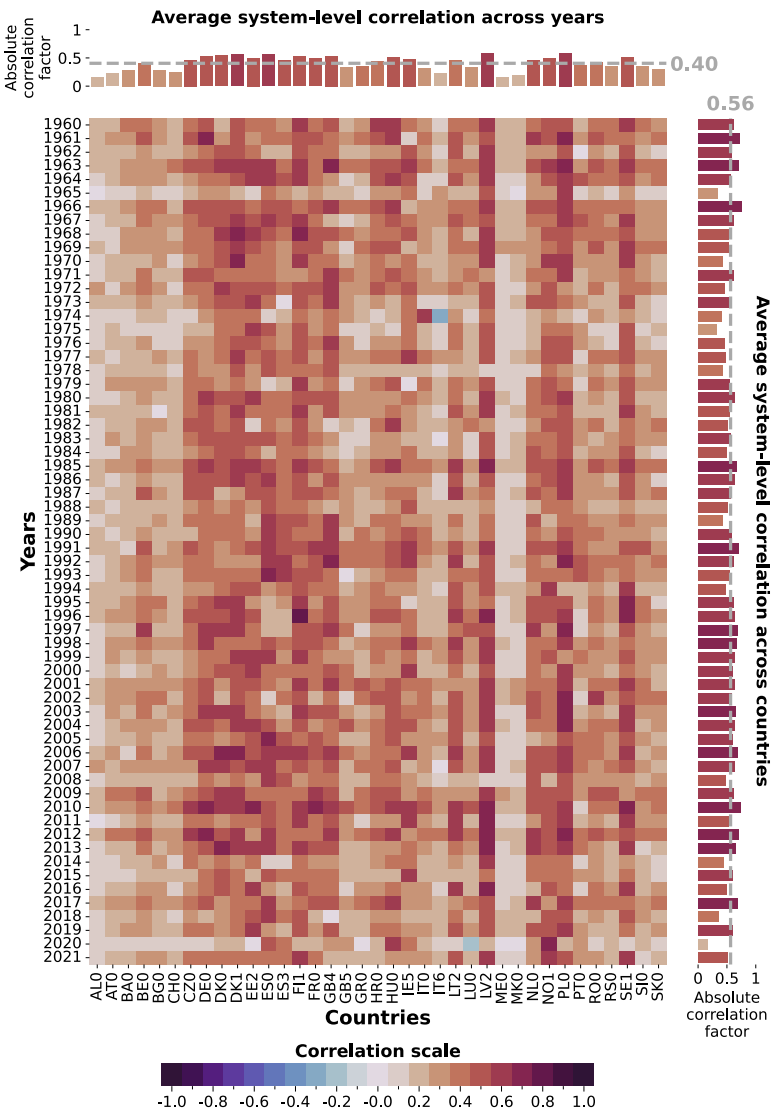
The overlap count reflects how many event pairs exhibit a temporal alignment. A high overlap count indicates that many predicted and benchmark events are overlapping.

- **Symmetric Accuracy:** Fraction of event pairs with non-zero temporal overlap that exhibit strong temporal alignment.

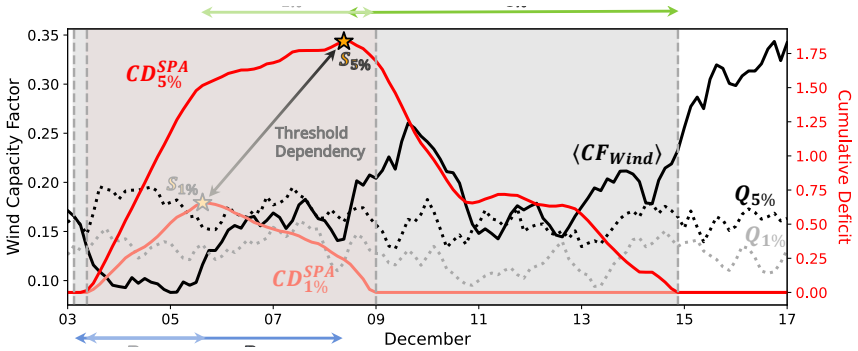
$$\boxed{\begin{aligned} \text{Accuracy} &= \left\langle \frac{\left|\left(\mathcal{TP}_{I,\tau^*} \cap \mathcal{TP}_{B,\tau^*}\right) \cup \mathcal{F}_{\tau^*}\right|}{|\mathcal{M}_{\tau^*}|} \right\rangle_{w_{\tau^*}}, \quad \text{with} \\ \mathcal{F}_{\tau^*} &= \left\{ (i,j) \in \mathcal{M}_{\tau^*} \;\middle|\; O_{ij,\tau^*} = |e^i_{I,\tau^*}| \;\vee\; O_{ij,\tau^*} = |e^j_{B,\tau^*}| \right\} \end{aligned}} \tag{28}$$

The symmetric accuracy denotes the fraction of event pairs whose events are either matched in both directions (i.e., counted as true positives for both predicted and benchmark sets) or are fully contained within each other. A high symmetric accuracy indicates that matched predicted and benchmark events are temporally well-aligned.

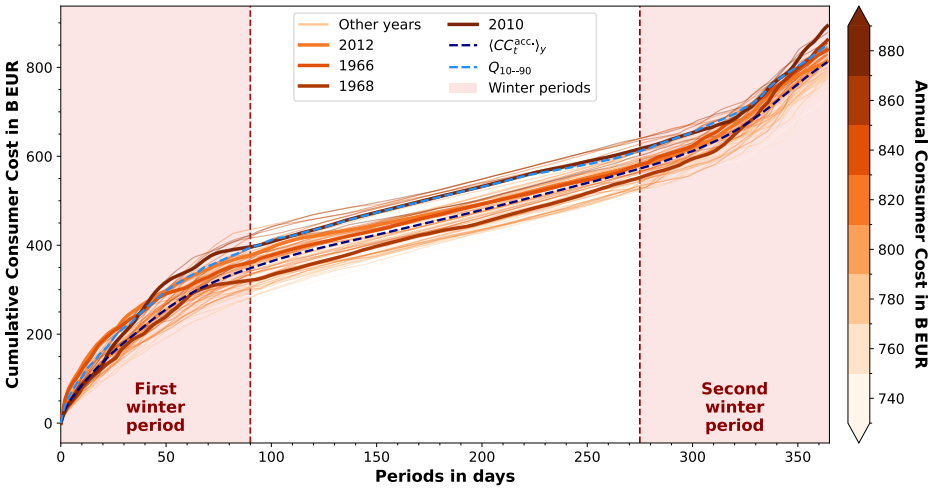
Supplementary Figures



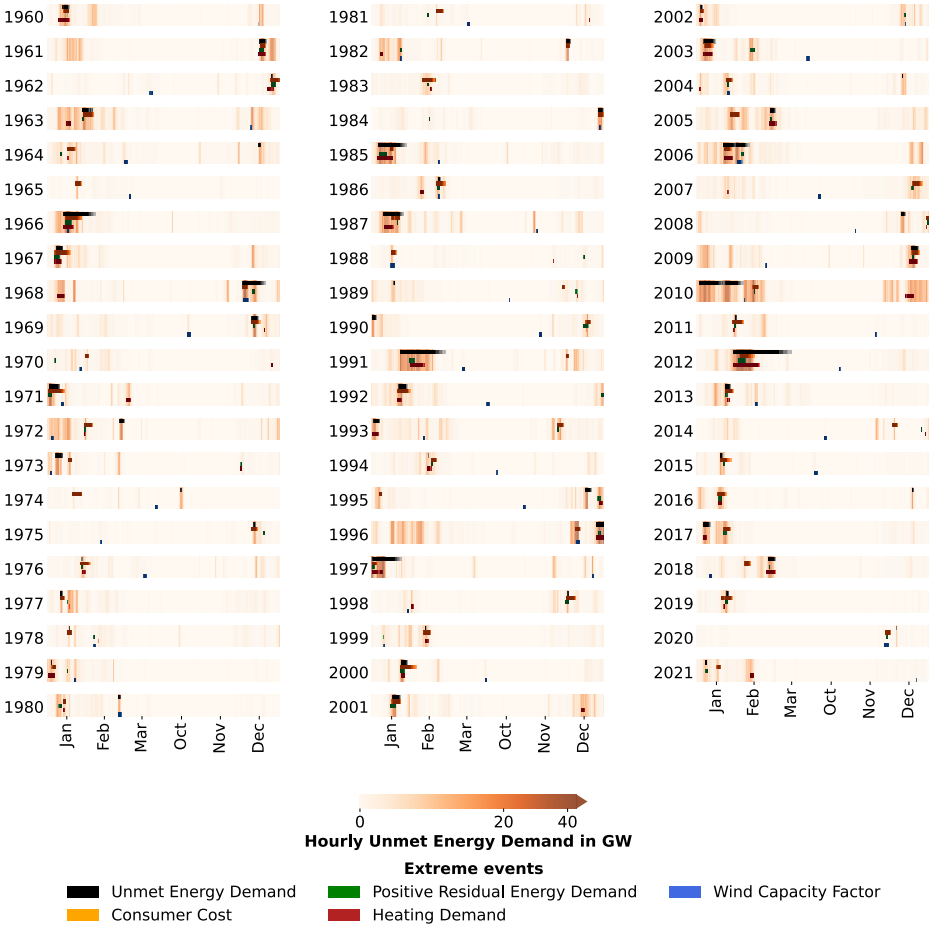
Supplementary Figure 1: **Correlation between *Positive Residual Energy Demand* and *Unmet Energy Demand* across all weather years (1960–2021) and individual countries.** Each cell shows the Pearson correlation for one country–year combination. Bar plots display averages across time (top, all years) and across space (right, all countries). Correlations are generally lower and more heterogeneous than for *Consumer Cost*, showing that *Positive Residual Energy Demand* only partially captures the spatio-temporal structure of physical system stress.



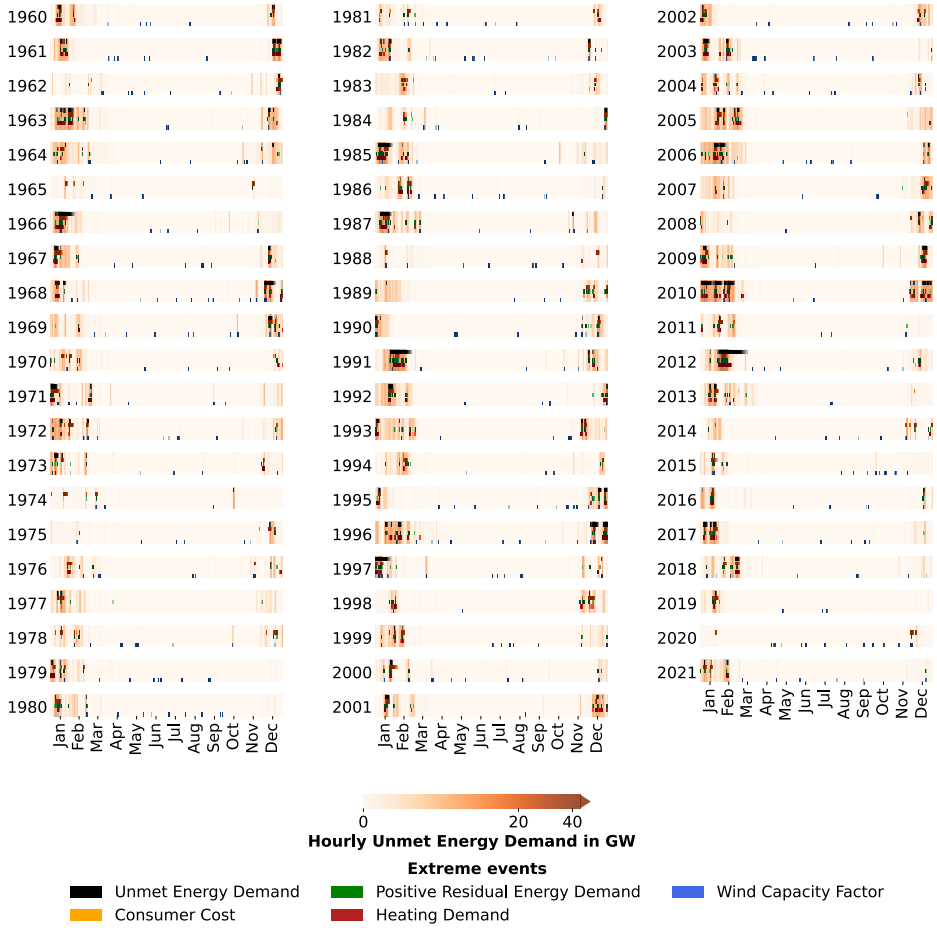
Supplementary Figure 2: **Exemplary extreme event identification for the *Wind Capacity Factor* in December 1968.** Shown is the time series of the *Wind Capacity Factor* CF_{Wind} (black) together with quantile thresholds $Q_{1\%}$ and $Q_{5\%}$ (dotted lines). The cumulative deficits CD_{τ}^{SPA} (shaded in red) quantify event severity. For this example, the threshold-specific build-up B_{τ}, recovery R_{τ}, and severity peaks S_{τ} are indicated for the chosen quantile thresholds $\tau \in \{1\,\%, 5\,\%\}$.



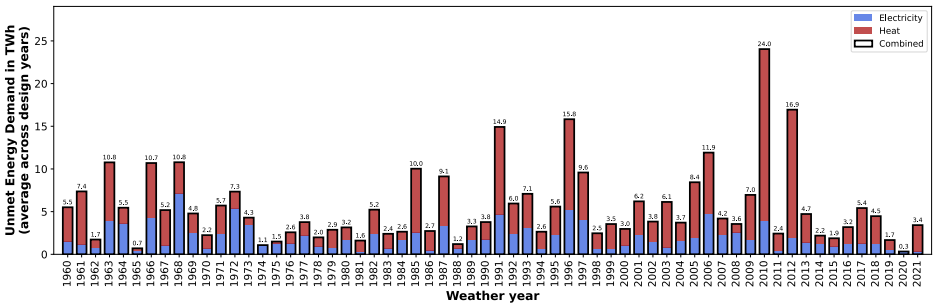
Supplementary Figure 3: **Cumulative *Consumer Cost* as a function of period length (days).** For each weather year, the curve shows the cumulative *Consumer Cost* of the most expensive periods up to the given length, ending at the annual total. Highlighted years (1966, 1968, 2010, 2012) represent weather-driven stress years with particularly high costs. Two phases of steep increase correspond to winter months (January–March and October–December), demonstrating seasonal clustering of stress.



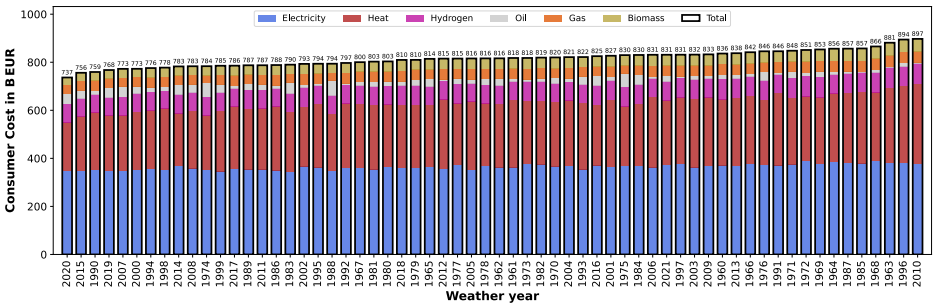
Supplementary Figure 4: Temporal alignment between the annual maximum extreme events identified by different indicators. Each row represents the winter months of a weather year (1960–2021). Background shading represents the *Unmet Energy Demand* (benchmark) timeseries. Colored bars show concurrent events for *Unmet Energy Demand* (black), *Consumer Cost* (orange), *Positive Residual Energy Demand* (green), *Heating Demand* (red), and *Wind Capacity Factor* (blue).



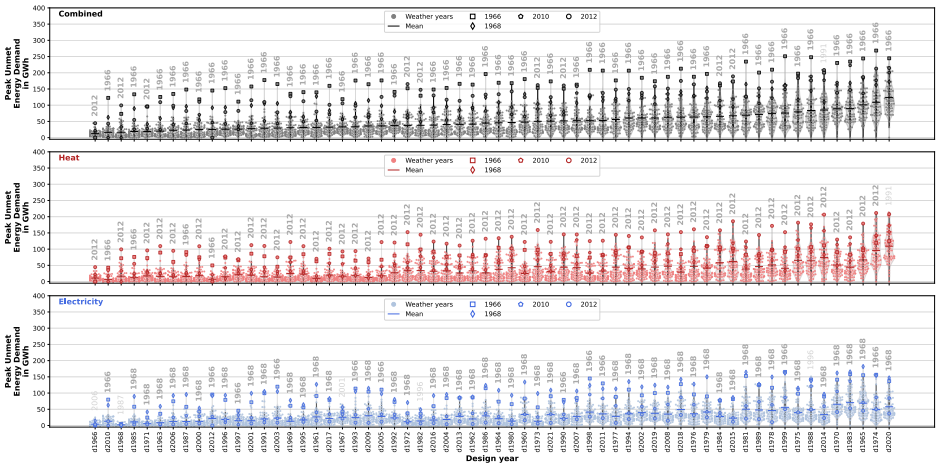
Supplementary Figure 5: Temporal alignment between all identified extreme events identified by different indicators. Each row represents the winter and summer months of a weather year (1960–2021). Background shading represents the *Unmet Energy Demand* (benchmark) timeseries. Colored bars show concurrent events for *Unmet Energy Demand* (black), *Consumer Cost* (orange), *Positive Residual Energy Demand* (green), *Heating Demand* (red), and *Wind Capacity Factor* (blue). The figure highlights systematic overprediction by meteorological indicators and frequent partial matches of benchmark events by *Positive Residual Energy Demand*.



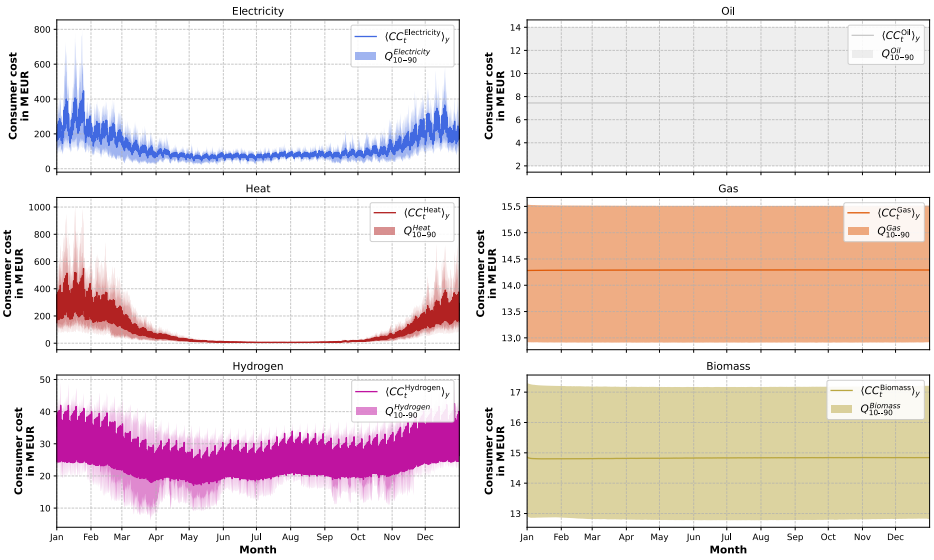
Supplementary Figure 6: **Annual *Unmet Energy Demand* for all weather years (1960–2021), averaged over all design years.** Bars show annual totals for electricity (blue), heat (red), and the combined *Unmet Energy Demand* (black outline). The figure highlights whether annual shortfalls are dominated by electricity or heating, with some years exhibiting particularly high combined *Unmet Energy Demand*.



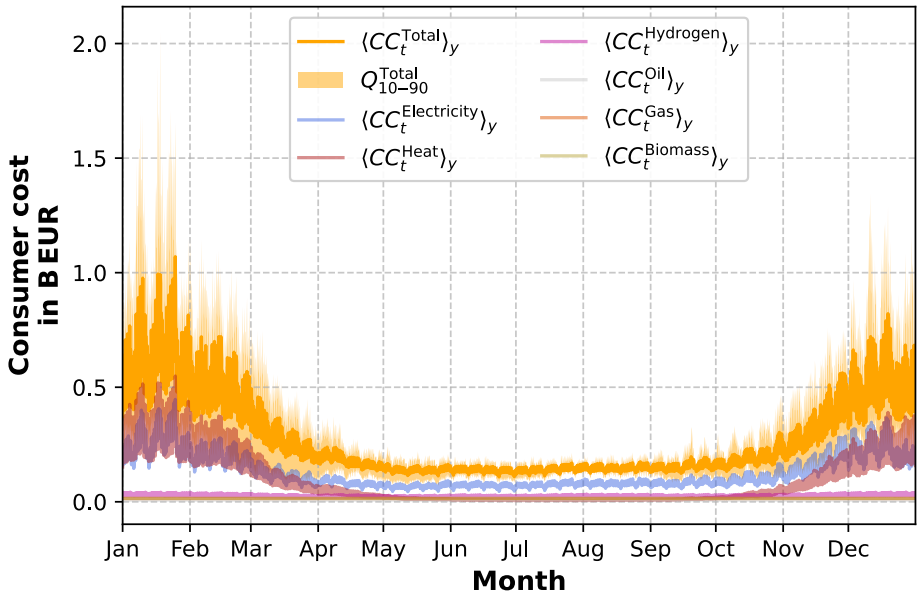
Supplementary Figure 7: **Annual *Consumer Cost* by energy carrier across all weather years (1960–2021).** Stacked bars show contributions from electricity, heating, hydrogen, oil, gas, and biomass, with totals outlined in black. The figure illustrates strong inter-annual variability in total *Consumer Cost* and the changing composition of consumer expenditures.



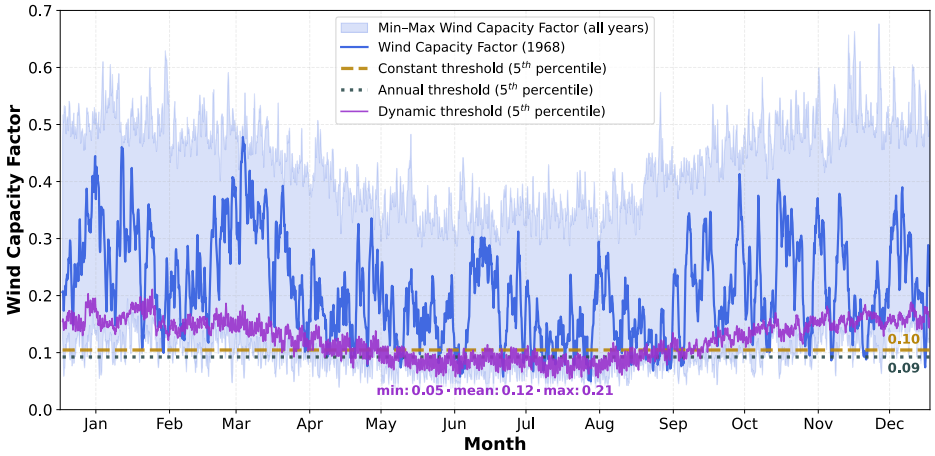
Supplementary Figure 8: Peak annual *Unmet Energy Demand* across weather years (1960–2021). The figure presents peak values for combined (top), heating (middle), and electricity (bottom) *Unmet Energy Demand*. Weather years are ordered by the average combined peak value across design years. Highlighted points indicate selected years (1966, 1968, 2010, 2012). The results underline the role of sector coupling, as combined peaks differ substantially from those in electricity or heating alone.



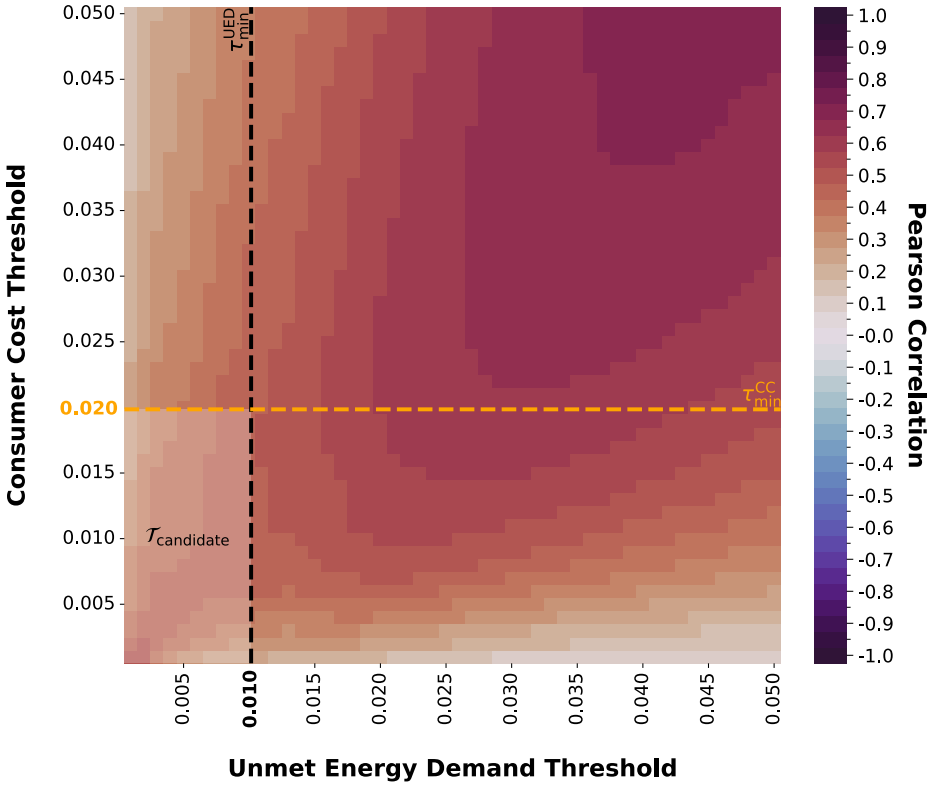
Supplementary Figure 9: **Time-resolved *Consumer Cost* (CC_t) separated by energy carrier ($x \in \mathcal{X}$).** Shown are mean values ($\langle CC_t^x \rangle_y$) and interquartile range (Q^x_{10-90}) across all design years. Electricity and heating display simultaneous winter peaks, indicating cross-sectoral system stress. Fossil oil and biomass costs remain nearly constant and largely weather-independent, since they can be stored and transported at no cost.



Supplementary Figure 10: **Time-resolved Consumer Cost (CC_t) by individual energy carrier ($x \in X$) and their aggregated total (CC_t^{Total}).** Shown are mean values ($\langle CC_t^x \rangle_y$ and $\langle CC_t^{Total} \rangle_y$) and the interquartile range (Q_{10-90}^{Total}) across all design years. The aggregate *Consumer Cost* peaks at around 2 B EUR, with significant increases in the winter months. Electricity and heating dominate the seasonal pattern, with pronounced winter peaks, whereas fossil oil and biomass remain nearly constant and weather-independent.



Supplementary Figure 11: **Illustration of threshold types for extreme event identification.** Example for the *Wind Capacity Factor* in 1968, showing constant, annual, and dynamic thresholds applied to the same time series. Constant thresholds use a fixed upper quantile across all years, annual thresholds recalculate the quantile each year, and dynamic thresholds adjust to the seasonal cycle using smoothed three-hourly quantiles. The figure illustrates how the different threshold types capture variability and seasonality.



Supplementary Figure 12: **Correlation-based threshold selection for SPA event time series.** Correlation matrix between *Consumer Cost* and *Unmet Energy Demand* SPA timeseries across threshold levels. Correlations increase at very low quantile thresholds, but these values lie far from the knee-point that marks the onset of extremal behavior. The region of statistically indistinguishable thresholds selected in our work ($\mathcal{T}^*$) is highlighted, while the ceiled knee-value thresholds are indicated by dashed lines.

Supplementary Tables

Supplementary Table 1: **Indicator-specific threshold definitions for extreme event identification.** Shown are the optimal percentile threshold ranges $\mathcal{T}^*$, the corresponding absolute values, and the applied threshold type (constant, annual, or dynamic) after indicator transformation, used in the extended Sequent Peak Algorithm (SPA).

Indicator	**Percentiles %**	**Absolutes**	**Threshold Type**
	lower · optimal · upper	lower · optimal · upper	
Unmet Energy Demand	99.0 · 99.4 · 99.6	13.5 · 19.2 · 24.6 GW	constant
Consumer Cost	99.0 · 99.2 · 99.4	121 · 133 · 150 Mio. EUR	annual
Positive Residual Energy Demand	99.0 · 99.5 · 99.7	569 · 664 · 723 GW	constant
Heating Demand	95.0 · 96.3 · 97.2	869 · 905 · 935 GW	constant
Wind Capacity Factor	98.8 · 99.1 · 99.4	-7.5 · -7.1 · -6.7 %	dynamic

Supplementary Table 2: Structural characteristics of identified extreme events across indicators. Reported values include minimum, mean, and maximum for frequency, duration, build-up, recovery, and coverage relative to annual *Consumer Cost* and annual *Unmet Energy Demand*. Results are given separately for each indicator.

Characteristic	Unmet Energy Demand	Consumer Cost	Positive Residual Energy Demand	Heating Demand	Wind Capacity Factor	Units
	Min · Mean · Max	Min · Mean · Max	Min · Mean · Max	Min · Mean · Max	Min · Mean · Max	
Frequency	0.0 · 3.5 · 21.0	1.0 · 3.7 · 9.1	1.6 · 12.0 · 27.6	0.0 · 6.1 · 34.8	0.0 · 6.4 · 13.7	#/year
Duration	1.7 · 45.6 · 921.8	1.7 · 43.1 · 261.2	1.7 · 12.4 · 152.3	1.7 · 20.4 · 458.1	1.7 · 20.5 · 88.5	h
Build-Up	0.6 · 22.0 · 566.7	0.6 · 21.0 · 193.5	0.6 · 4.2 · 99.0	0.6 · 8.8 · 179.7	0.6 · 7.1 · 42.5	h
Recovery	1.1 · 23.6 · 366.0	1.1 · 22.1 · 119.8	1.1 · 8.2 · 112.7	1.1 · 11.6 · 330.5	1.1 · 13.4 · 68.9	h
Consumer Cost	0.2 · 9.1 · 39.2	7.2 · 9.6 · 14.2	1.0 · 8.7 · 21.4	0.1 · 7.1 · 23.9	0.1 · 3.2 · 10.6	%/year
Coverage	0.0 · 2.0 · 28.2	0.1 · 2.2 · 14.2	0.0 · 0.6 · 11.6	0.0 · 0.9 · 21.8	0.0 · 0.4 · 5.0	%/e
Unmet Energy	2.1 · 39.3 · 86.8	15.6 · 38.6 · 65.8	0.6 · 32.7 · 76.9	0.2 · 27.3 · 86.9	0.0 · 8.3 · 39.6	%/year
Demand Coverage	0.1 · 8.7 · 86.1	0.0 · 8.9 · 65.8	0.0 · 2.3 · 53.3	0.0 · 3.5 · 84.9	0.0 · 1.1 · 34.5	%/e
	Winter · Total	Winter · Total	Winter · Total	Winter · Total	Winter · Total	
Count	216 · 216	229 · 229	741 · 742	377 · 378	184 · 397	#

Supplementary Table 3: **Evaluation metrics for each stress indicator using event-based identification.** Metrics include precision, recall, F1-score, and temporal overlap, each reported as minimum, mean, and maximum values. Results are shown separately for timestep matching, event matching, and maximum event matching.

Indicator	Precision			Recall			F1-Score			Overlap		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Timestep Matching												
Consumer Cost	26.6	43.6	56.5	32.4	48.9	71.1	38.7	44.7	48.6	32.4	48.9	71.1
Positive Residual Energy Demand	25.2	44.4	68.2	19.9	40.0	61.7	30.8	40.2	42.8	19.9	40.0	61.7
Heating Demand	9.4	21.6	34.0	69.7	82.0	89.0	16.9	33.6	45.7	69.7	82.0	89.0
Wind Capacity Factor	6.0	9.1	11.1	5.4	9.4	15.6	7.3	9.0	10.2	5.4	9.4	15.6
Event Matching												
Consumer Cost	21.3	43.2	53.8	33.2	49.8	71.4	32.8	44.8	48.8	56.6	68.0	85.2
Positive Residual Energy Demand	18.0	38.1	65.6	19.8	42.7	73.9	29.0	37.6	42.0	24.3	35.4	58.9
Heating Demand	2.0	6.2	11.2	73.5	83.6	90.1	3.8	11.5	19.5	81.4	87.4	95.9
Wind Capacity Factor	5.2	7.8	9.6	6.2	10.4	16.1	6.5	8.8	10.3	39.0	45.2	56.9
Maximum Event Matching												
Consumer Cost	16.1	35.0	41.9	48.2	53.8	68.3	26.1	41.7	45.6	70.2	83.6	94.9
Positive Residual Energy Demand	24.2	37.0	42.4	7.1	20.9	36.6	12.2	25.6	35.4	16.1	37.9	65.9
Heating Demand	6.5	22.9	36.1	50.9	53.3	57.1	11.5	31.2	43.1	88.5	97.5	100.0
Wind Capacity Factor	4.9	8.3	11.7	1.8	4.8	12.2	2.9	5.7	9.9	30.4	40.8	60.3
Units	%			%			%			%		

Supplementary Table 4: **Parameter settings for automated threshold selection.** The table specifies the bootstrap setup, candidate threshold range, and statistical tests used to identify robust and statistically indistinguishable thresholds.

Parameter	Value	Description
$CI_{1\sigma}$	[0.16, 0.84]	Bootstrap confidence interval
Δ_τ	0.001	Increments between candidate thresholds
k	2500	Number of bootstrap replicates
m	500	Number of quantile levels per bootstrap replicate
$\mathcal{N}_e(\tau)$	62	Minimum number of exceedances (distinct extreme events)
$\overline{p}_{\mathrm{AD}}(\tau)$	0.05	Anderson–Darling test significance level
$q_{\min}$	95^{th}	Minimum admissible percentile for candidate thresholds