GEFC 2012: Electricity Load Forecasting and Causal Impact Analysis

1 Data Preprocessing

The following preprocessing steps were applied to the raw datasets:

- **Zone Aggregation**: Created a new "Zone 21" by summing the hourly electricity usage of the first 20 zones.
- **Data Reshaping**: Converted both electricity and temperature datasets from 24-column "Wide" daily rows into single-column "Long" hourly rows.
- **Timestamping**: Combined year, month, day, and hour columns into a single `ds` (datestamp) object.
- **Weighting**: Assigned Zone 21 a 20x higher importance weight than individual zones, as specified in `weights.csv`.
- **Station Mapping**: Reorganized temperature data so each of the 11 weather stations has its own unique column.
- **Temperature mapping**: performed a correlation analysis to identify and assign the single weather station whose temperature profile most closely matched the historical load fluctuations of each specific zone.
- **Temporal Features**: Added `is_holiday`, `is_weekend`, and `day_of_week`.

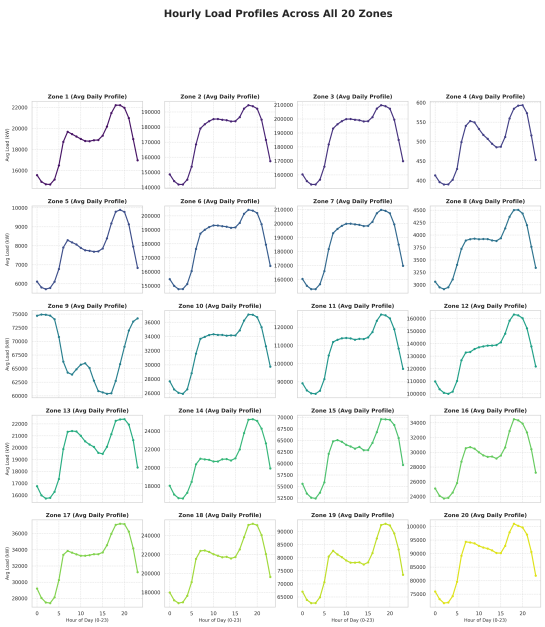
	zone_id	ds	load	weight	temp_s1	temp_s2	temp_s3	\
0	1	2004-01-01 00:00:00	16853.0	1.0	46.0	38.0	44.0	
1	1	2004-01-01 01:00:00	16450.0	1.0	46.0	36.0	42.0	
2	1	2004-01-01 02:00:00	16517.0	1.0	45.0	35.0	40.0	
3	1	2004-01-01 03:00:00	16873.0	1.0	41.0	30.0	36.0	
4	1	2004-01-01 04:00:00	17064.0	1.0	39.0	30.0	34.0	

	temp_s4	temp_s5	temp_s6	temp_s7	temp_s8	temp_s9	temp_s10	temp_s11	\
0	45.0	42.0	44.0	45.0	43.0	41.0	42.0	36.0	
1	43.0	42.0	43.0	44.0	44.0	39.0	43.0	32.0	
2	41.0	40.0	42.0	41.0	42.0	36.0	43.0	31.0	
3	37.0	39.0	38.0	40.0	34.0	35.0	39.0	30.0	
4	33.0	40.0	38.0	35.0	30.0	33.0	35.0	34.0	

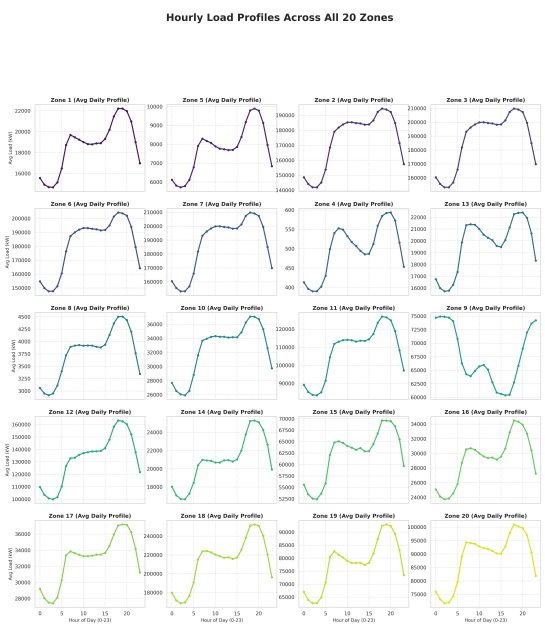
	mapped_station_id	mapped_temp
0	2	38.0
1	2	36.0
2	2	35.0
3	2	30.0
4	2	30.0

Figure 1: Preprocessed Data

2 Exploratory Data Analysis (EDA)



(a) Hourly Load Average



(b) Hourly Load Average (arranged)

Figure 2: Hourly Load Averages

This shows importance of hour column.

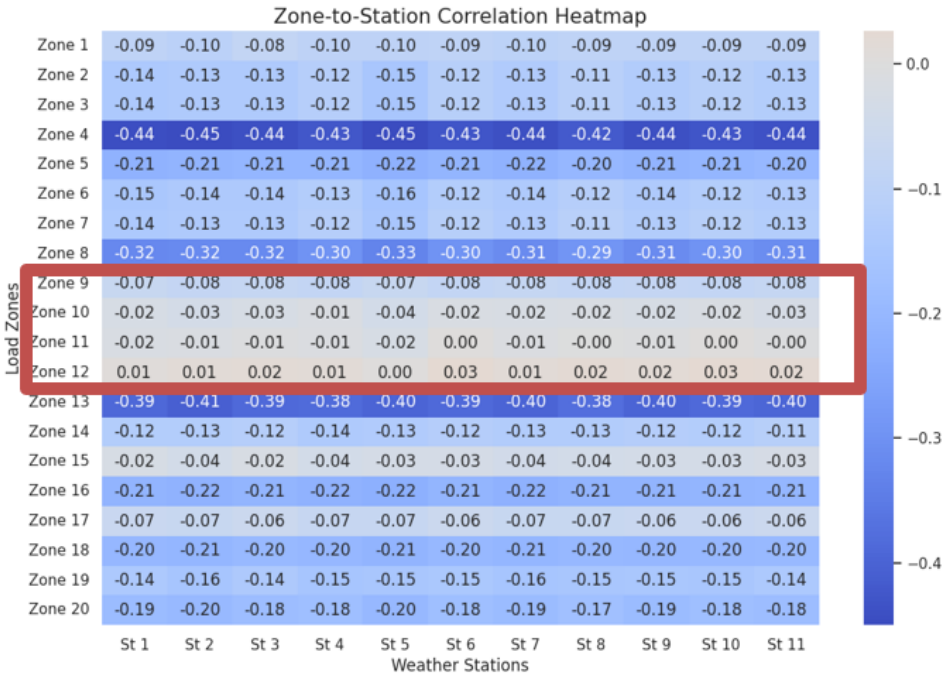


Figure 3: Zone to station correlation Heatmap

Figure 4: Weekly Load Distribution

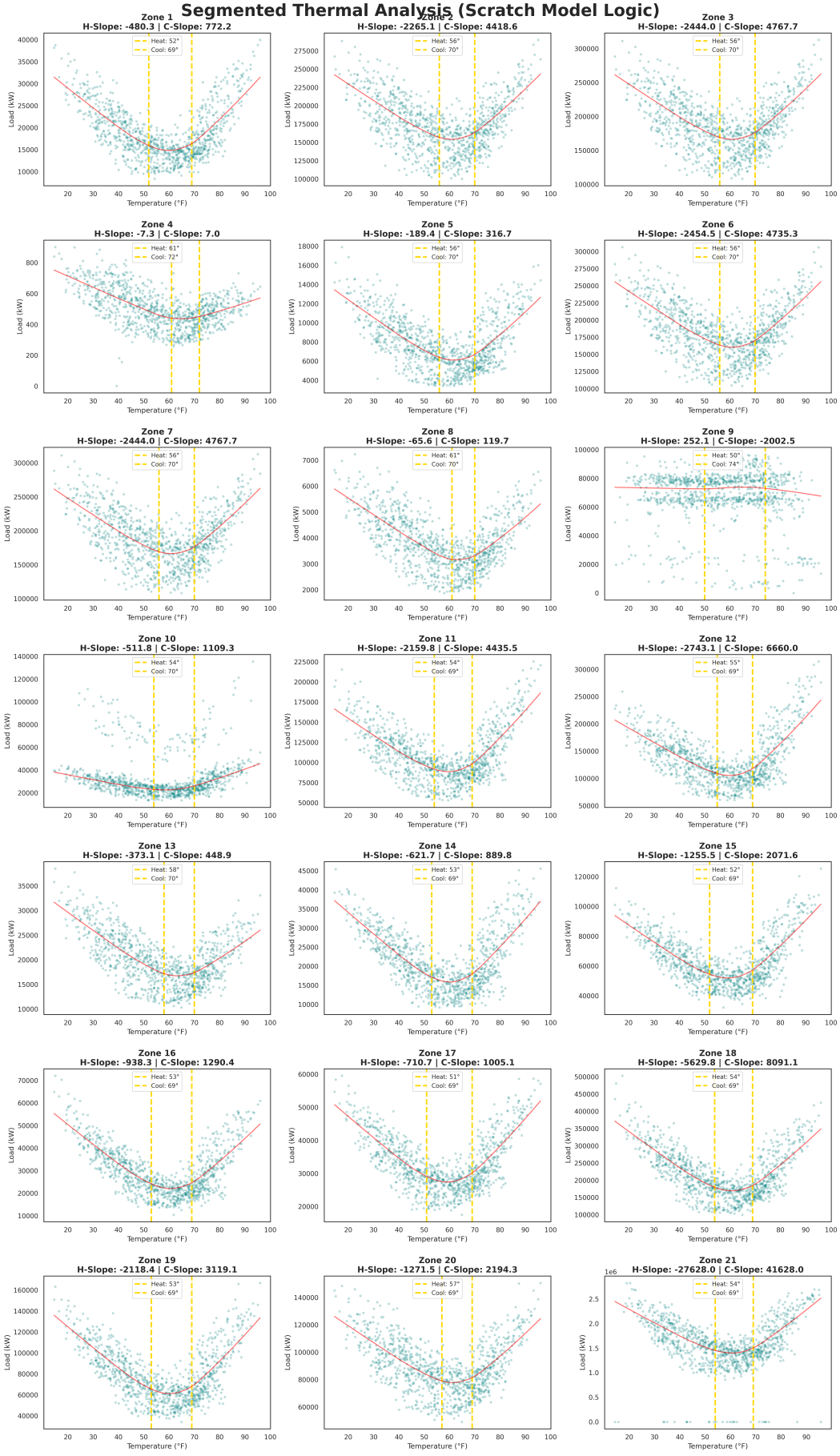


Figure 5: Thermal Analysis

All_Zones_Monthly_Thermal_Signatures The EDA revealed several key insights:

- **Hourly Dependency**: Current load is heavily dependent on the **hour** of the day, following a clear daily cycle.
- **Zone Consistency**: Patterns across zones are similar (except Zone 9), allowing for a unified model structure across the region.
- **Weekly Seasonality**: Load fluctuates significantly based on the **day of the week**, distinguishing between business days and weekends.
- **Temperature Correlation**: Analysis showed a surprisingly **low correlation** between individual zone loads and their specifically mapped weather stations.
- **Station Averaging**: To reduce localized noise and capture the regional thermal trend, it was found more effective to use the **average temperature of all 11 stations** as the primary predictor rather than single-station data.

3 Thermal Signatures and Piecewise Linear Regression

To capture the non-linear relationship between temperature and electricity demand, a **Segmented Thermal Analysis** was performed for each zone. Rather than assuming global thresholds, T_{heat} and T_{cool} were calculated by minimizing the Sum of Squared Errors (SSE) across a three-segment model:

1. **Heating Segment**: A linear fit for temperatures below T_{heat}.
2. **Comfort Segment**: A flat baseline representing the mean load when no HVAC(Heating, Ventilation, and Air Conditioning) is required.
3. **Cooling Segment**: A linear fit for temperatures above T_{cool}.

The optimization problem for each zone is defined as:

$$\min_{T_{\text{heat}}, T_{\text{cool}}} \left(SSE_{\text{heating}} + SSE_{\text{comfort}} + SSE_{\text{cooling}} \right) \tag{1}$$

Subject to the constraint $T_{\text{heat}} < T_{\text{cool}}$. The resulting slopes ($H_{\text{slope}}$ and C_{slope}) represent the sensitivity of the load (kW/°F) to temperature changes in their respective regimes.

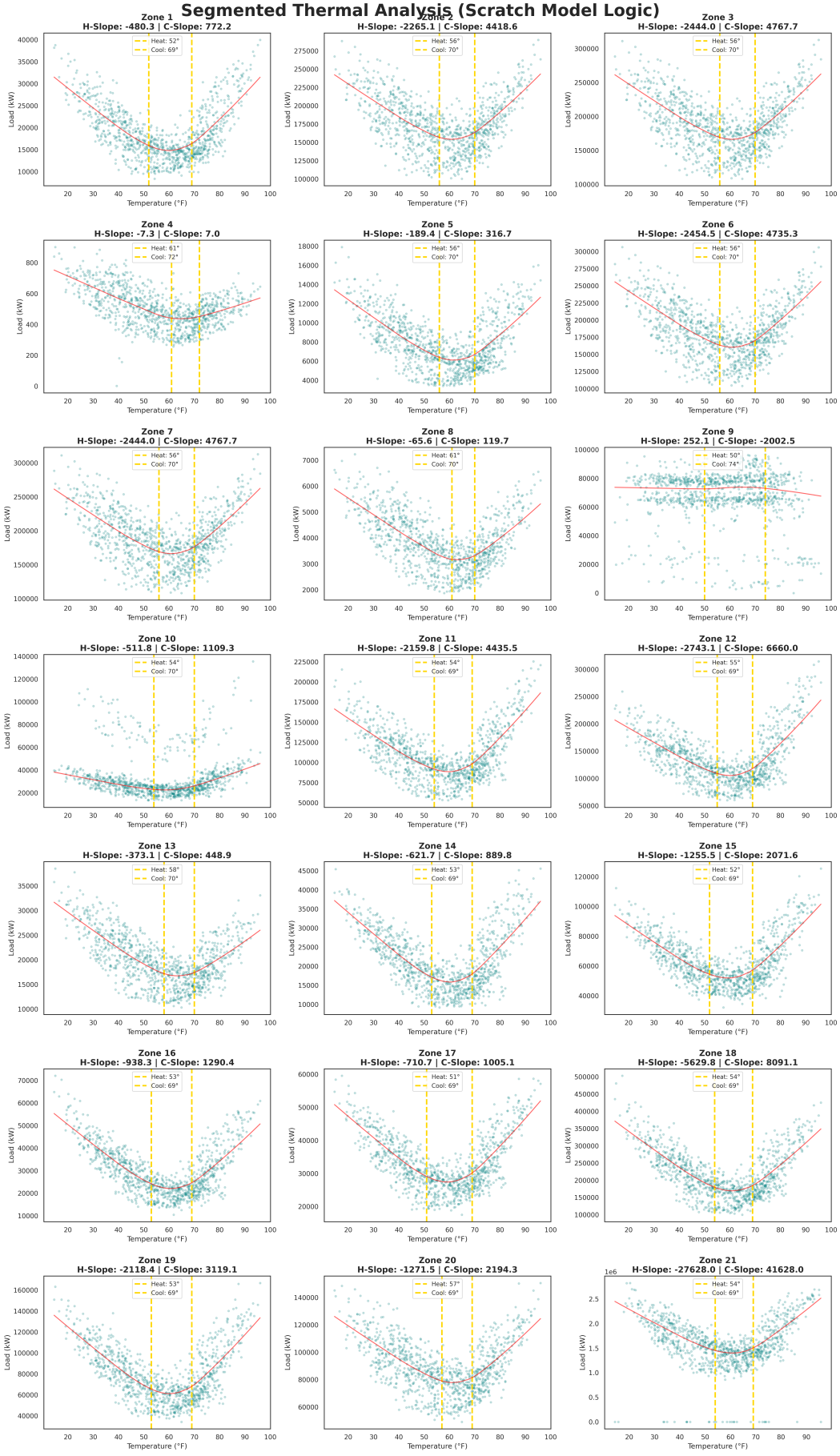


Figure 6: Segmented Thermal Analysis

These calculated "elbows" allow for the creation of high-fidelity Heating Degree Hours (HDH) and Cooling Degree Hours (CDH) features tailored to the unique infrastructure of each specific zone.

How T_{heat}, T_{cool}, and the slopes are calculated:

The method finds the two key temperature points—the heating threshold (T_{heat}) and the cooling threshold (T_{cool})—that best explain how electricity demand changes with outdoor temperature.

This is done by trying many possible pairs of thresholds and choosing the pair that gives the smallest total error when fitting three straight line segments to the data:

1. **Below T_{heat}**: Demand rises quickly as it gets colder (heating turns on). The best straight line through these points gives the **heating slope** (H_{slope}) in kW per °F.

2. **Between T_{heat} and T_{cool}**: Demand is fairly flat. We use the average demand here (a horizontal line).

3. **Above T_{cool}**: Demand rises quickly as it gets hotter (cooling turns on). The best straight line through these points gives the **cooling slope** (C_{slope}) in kW per °F.

The program searches for the values of T_{heat} and T_{cool} (with $T_{\text{heat}} < T_{\text{cool}}$) that minimise the sum of squared differences between the actual data and the three-segment model.

With these zone-specific thresholds and slopes, we can create accurate Heating Degree Hours (HDH) and Cooling Degree Hours (CDH) that match how buildings in this particular area actually respond to cold and hot weather, instead of using a fixed reference temperature for every zone.

4 Thermal Signatures and Momentum Features

The load was found to be highly dependent on Cooling Degree Hours (CDH) and Heating Degree Hours (HDH):

$$CDH = \max(0, T_{\text{actual}} - T_{\text{cool}}) \tag{2}$$

$$HDH = \max(0, T_{\text{heat}} - T_{\text{actual}}) \tag{3}$$

Lag features were created to capture different cycles:

- **load_lag_1h (Momentum)**: Predicts current load based on the state 60 minutes prior.
- **load_lag_24h (Daily Rhythm)**: Uses yesterday's behavior at the same time as a blueprint.
- **load_lag_168h (Weekly Pattern)**: Captures the 7-day cycle to differentiate workdays from weekends.

5 Predictive Model

A separate Random Forest model (100 estimators, max depth of 10) was trained for each individual zone. The model utilizes a feature set consisting of *cdh*, *hdh*, *hour*, *dayofweek*,

month, `load_lag_1h`, `load_lag_24h`, and `load_lag_168h`. Training was conducted on data up to $max(ds) - 8$ weeks, with the subsequent 8 weeks reserved for testing.

To assess performance, the Weighted Mean Absolute Percentage Error (WMAPE) was calculated as follows:

$$WMAPE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{\sum_{t=1}^{n} |y_t|} \times 100 \tag{4}$$

Results

Zone	MAE (kW)	WMAPE (%)	Zone	MAE (kW)	WMAPE (%)
Zone 1	453.17	2.51	Zone 12	3193.66	2.37
Zone 2	2644.19	1.58	Zone 13	705.36	4.16
Zone 3	2853.79	1.59	Zone 14	925.68	4.59
Zone 4	15.72	3.62	Zone 15	2315.42	3.92
Zone 5	225.96	3.38	Zone 16	963.01	3.61
Zone 6	2727.44	1.57	Zone 17	814.67	2.44
Zone 7	2853.79	1.59	Zone 18	6246.80	3.06
Zone 8	100.29	2.93	Zone 19	3078.29	4.14
Zone 9	4063.72	6.17	Zone 20	2351.67	2.72
Zone 10	2907.60	3.83	Zone 21	31383.77	1.92
Zone 11	2206.71	2.01	**OVERALL**	**3477.65**	**3.03**

6 Causal Impact Analysis

To isolate the true physical impact of weather and holidays on grid load, we implement a framework based on Invariant Causal Prediction (ICP) and Structural De-confounding. Unlike standard predictive models, this approach identifies parameters that are invariant across different thermal regimes.

Problem Formulation

The is modeled system using a Structural Equation Model (SEM) defined by the variables $\mathcal{V} = \{L, T, C, \epsilon\}$:

- **Outcome (L):**load.
- **Treatments (T):**$[cdh, hdh, is\_holiday]$.
- **Confounders (C):**$[hour, dayofweek, month, is\_weekend, load\_lag_{1h,24h,168h}]$.

Structural De-confounding (Purification)

To eliminate the "back-door" influence of time-based patterns ($T \leftarrow C \rightarrow L$), we project both the load and treatments into the subspace orthogonal to the confounder matrix C. The purified anomalies, L' and T', are calculated as:

$$L' = (I - P_C)L, \quad T' = (I - P_C)T \tag{5}$$

where $P_C = C(C^\top C)^{-1}C^\top$ is the orthogonal projection matrix onto the column space of C.

Invariant Causal Estimation

We seek the invariant parameter vector γ^* that satisfies the purified structural equation:

$$L' = \gamma^* T' + \epsilon_{L'} \tag{6}$$

The parameter γ^* is estimated separately for each *zone* to account for geographical heterogeneity in grid infrastructure.

Evaluation of Global Invariance

The invariance of the structural parameter γ^* was evaluated by testing the Null Hypothesis $H_0 : E[\epsilon|E = e_1] = E[\epsilon|E = e_2] = E[\epsilon|E = e_3]$, where $\{e_1, e_2, e_3\}$ represent the Cold, Mild, and Hot thermal regimes respectively.

This was implemented using a one-way Analysis of Variance (ANOVA) F-test on the purified residuals:

$$r = L' - \gamma^* T' \tag{7}$$

Results

Zone	CDH	HDH	Holiday	P-Value
Zone 1	158.09	90.27	48.32	0.000000000
Zone 2	589.09	250.42	-2283.93	0.000624589
Zone 3	635.63	270.20	-2464.37	0.000624617
Zone 4	1.46	0.68	7.05	0.000001315
Zone 5	65.74	31.93	59.55	0.004722163
Zone 6	661.16	285.01	-2342.21	0.000720401
Zone 7	635.63	270.20	-2464.37	0.000624617
Zone 8	17.63	6.36	25.39	0.041638393
Zone 9	-299.96	-56.70	364.49	0.003106494
Zone 10	100.67	41.40	-354.37	0.012969885
Zone 11	757.41	363.20	-597.16	0.000010220
Zone 12	1219.95	504.53	591.20	0.000000000
Zone 13	83.38	56.22	86.54	0.132065956
Zone 14	231.14	141.49	202.38	0.309439682
Zone 15	486.13	291.28	-62.41	0.028571672
Zone 16	281.20	184.72	225.03	0.000000000
Zone 17	248.40	161.48	-127.18	0.806170160
Zone 18	1821.60	1155.51	232.61	0.006388326
Zone 19	765.73	490.16	296.41	0.001628175
Zone 20	349.57	189.74	-419.09	0.075573367
Zone 21	4842.78	1853.45	2943.38	0.000000000

Table 1: Causal Coefficients and Invariance P-Values by Zone