# Causal Feature Selection for Weather-Driven Residential Load Forecasting

Elise Zhang*, François Mirallès†, Stéphane Dellacherie‡§, Di Wu*, Benoit Boulet*
*Department of Electrical and Computer Engineering, McGill University, Montréal, QC, Canada
†Hydro-Québec Research Institute, Varennes, QC, Canada
‡Energy System Control Planning Division, Hydro-Québec, Montréal, QC, Canada
§Department of Computer Science, Université du Québec à Montréal (UQÀM), Montréal, QC, Canada
Emails: elise.zhang@mail.mcgill.ca, {miralles.francois, dellacherie.stephane}@hydroquebec.com, {di.wu5, benoit.boulet}@mcgill.ca

***Abstract*—Weather is a dominant external driver of residential electricity demand, but adding many meteorological covariates can inflate model complexity and may even impair accuracy. Selecting appropriate exogenous features is non-trivial and calls for a principled selection framework, given the direct operational implications for day-to-day planning and reliability. This work investigates whether causal feature selection can retain the most informative weather drivers while improving parsimony and robustness for short-term load forecasting. We present a case study on Southern Ontario with two open-source datasets: (i) Independent Electricity System Operator (IESO)'s hourly consumption by Forward Sortation Areas; (ii) ERA5 weather reanalysis data. We compare different feature selection regimes (no selection, non-causal selection, PCMCI-causal selection) on city-level forecasting with well-established load forecasting models. We find that non-causal selection prioritizes radiation and moisture variables that show correlational dependence, whereas PCMCI-causal selection emphasizes more direct thermal drivers and prunes the indirect covariates. We detail the evaluation pipeline and report diagnostics on accuracy and extreme-weather robustness, positioning causal feature selection as a practical complement to modern forecasters when integrating weather into residential load forecasting.**

***Index Terms*—Causality, causal feature selection, load forecasting, time series analysis**

## I. Introduction

Short-term load forecasting (STLF) estimates near-future electricity demand (hours to days ahead) so operators can schedule generation, ensure reliability, and manage markets with minimal cost and risk. As grid-scale storage remains limited and costly, supply must closely track demand in real time; accurate STLF is therefore central to reliable and economical grid operations. At city level, STLF typically augments autoregressive load history with calendar signals (hour, weekday, holidays) and exogenous weather (temperature, humidity, cloud cover, radiative fluxes, precipitation). Prior studies [1], [2] find that appropriate weather integration improves forecasting performance. Yet, adding too many meteorological covariates inflates the feature space, which adds to model complexity and confounds the model with spurious correlations. This might negatively impact model performance, especially under seasonal regime shifts.

This paper studies whether **causal feature selection** (causal FS) helps isolate the most relevant weather drivers for electric load while reducing input dimensionality. We present a focused **case study on Southern Ontario**, pairing administrative-level residential electricity consumption from IESO *Hourly Consumption by FSA* reports with ERA5 [3] hourly meteorological reanalysis data. We compare no selection, non-causal filter, and PCMCI-based causal selection on single-city forecasting with SARIMAX [4], Random Forest (RF) Regression [5], GRU [6], TCN [7], and PatchTST [8].

Our contributions are summarized as follows: (*i*) A weather-informed load forecasting case study and evaluation pipeline that compares feature-selection regimes across forecasts on multiple cities; (*ii*) A feature-level analysis showing that causal FS favors direct thermal drivers consistent with meteorological mechanisms, whereas non-causal filtering retains more indirect correlates; (*iii*) Diagnostics on model accuracy, noise-sensitivity and robustness under extreme weather, positioning causal FS as a model-agnostic module for weather integration.

The remainder of the paper is structured as follows: Section II introduces related work; Section III formulates the problem; Section IV presents the causal and non-causal selection modules and evaluation design; Section V presents the experimental results; and Section VI summarizes key takeaways and briefly discusses future work.

## II. Related Work

### A. Feature Selection for Load Forecasting

STLF methods commonly combine load history with calendar features, and exogenous meteorology, reflecting the well-documented sensitivity of electricity demand to weather across seasons. Existing studies [1], [2] consistently report that weather is an important exogenous driver of electric load, and that appropriate weather integration improves accuracy and reliability. As the feature space grows with exogenous drivers, selecting appropriate input features becomes critical for both model generalization and efficiency. Controlling the feature set complexity also lowers computation time and memory usage, helping the forecastor meet the strict runtime constraint in production. Classical feature selection (FS) seeks to identify the minimal feature set that still exhibits optimal prediction performance [9]. FS approaches are often grouped into: (i) *filters* (e.g., maximal relevance/minimal redundancy, mRMR, using mutual information) [10]; (ii) *wrappers*, which search subsets with a predictive model in the loop [11]; and (iii) *embedded* methods that shrink/select during training (e.g., LASSO/Elastic Net) [12]. While these approaches can reduce

979-8-3315-8138-1

error and complexity, they primarily exploit correlations with the prediction target rather than underlying mechanisms (true causal relations), risking spurious selections and brittleness under distribution shift. Power systems studies also report that using all features is rarely optimal, reinforcing the need for principled selection [13].

### B. Causality and Causal Feature Selection

Causality captures the true data-generating mechanisms of systems, not mere predictability. Causal FS aims to select features by formal causal analysis [14] and seek variables that are **causally sufficient** to explain the target variable, rather than merely correlated with it. Under the Causal Markov and the Faithfulness Assumptions [15], [16], a causal Directed Acyclic Graph (DAG) entails that each node $X$ is conditionally independent of all its non-descendants, given its direct parents. In other words, the minimal subset of features that renders $X$ independent of all remaining variables contains features with a direct link with $X$ (parents, children, and "spouses"). This subset is referred to as the **Markov blanket** (MB) of $X$ (Fig. 1), and will be sufficient to explain $X$. Any feature outside of MB is redundant for predicting $X$, given MB. Consequently, selecting MB provides a minimal and causally sufficient feature set, aligning with the goal of FS. Classical work applies this principle directly to FS: KS algorithm [17] using MB-based screening, was established as a criterion for optimal feature selection as early as 1996; subsequent work on MB induction (e.g., IAMB [18]) realized the discovery of MBs around a target at scale. For **multivariate time series**, the relevant blanket envelops direct causal parents, both lagged and instantaneous. Time series causal discovery (e.g., TiMINo [19], PCMCI/PCMCI+ [20], [21]) target exactly the identification of such parents, offering causality-aware lens for FS tailored to time series-related tasks.

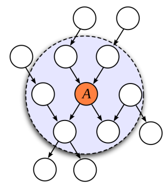

Fig. 1. Markov Blanket: Parents, children, spouses suffice for prediction under *Causal Markov + Faithfulness*

## III. Problem Formulation

We consider hourly residential electricity demand for a set of geographically proximate cities/regions $\mathcal{C}$. For each city/region $c \in \mathcal{C}$, we observe historical load $\mathbf{Y}_{0:t}^{(c)} = [y_0^{(c)}, y_1^{(c)}, y_2^{(c)}, ..., y_t^{(c)}]$, and extract datetime features $\mathbf{D}_{0:t}$ (e.g., hour-of-day, day-of-week, month-of-year, day-of-year, holidays), as well as corresponding meteorological predictors $\mathbf{W}_{0:t}^{(c)}$ (e.g., temperature, cloud, radiation, precipitation, etc.). Here, weather is treated as potential driver of electricity demand and will subject to feature selection as a preprocessing step.

### A. City-Level Short-Term Load Forecasting

Let $\mathcal{P}^{(c)} = [\mathbf{D}, \mathbf{Y}^{(c)}, \mathbf{W}^{(c)}]$ denote the full predictor set. Given a lookback window $L$ and forecast horizon $H$, we pose forecasting as supervised learning from historical predictors to future demand. In this study we use *one week* lookback and *one day* horizon, i.e., $L$=168 hours and $H$=24 hours. For city $c$, define its predictor history $\mathcal{P}_{\text{hist}}^{(c)} = [\mathbf{D}_{(t-L+1):t}, \mathbf{Y}_{(t-L+1):t}^{(c)}, \mathbf{W}_{(t-L+1):t}^{(c)}]$ and target $\mathbf{Y}_{\text{tgt}}^{(c)} = \mathbf{Y}_{(t+1):(t+H)}^{(c)}$. The task is, for each $c$, to learn a function $f^{(c)} : \mathcal{P}_{\text{hist}}^{(c)} \mapsto \mathbf{Y}_{\text{tgt}}^{(c)}$ that predicts future load values.

### B. Feature Selection

A central question in this case study is whether **causal** feature selection, as a principled selection method, improves forecasting quality and robustness, relative to no-selection and non-causal selection settings. Our working hypotheses are: **H1 (Parsimony without loss):** Causally selected feature subsets $\mathbf{P}_{hist}^{(\mathcal{C})}$ yield similar or lower error with reduced input dimensionality; **H2 (Robustness):** Models trained on causal features are more robust under extreme weather conditions.

## IV. Methodology

### A. PCMCI Causal Discovery for Feature Selection

PCMCI (Peter–Clark *PC algorithm* with Momentary Conditional Independence) [20] is a causal discovery method tailored to highly-interdependent multivariate time series. PCMCI improves upon the classic PC algorithm [16], [22] via a two-phase framework: **PC-Style Condition Selection**: Starting from a fully connected time-lagged graph with lag $0 \leq \tau \leq \tau_{max}$ (we set $\tau_{max} = 5$ to capture short-term weather impact while reducing computational complexity by keeping the search space tractable), PCMCI iteratively prunes some links through PC-style conditional independence tests. The output is a candidate set of time-lagged causal parents $Pa(X_t^{(j)})$ for $X_t^{(j)}$, variable $j$ at timestamp $t$. **Momentary Conditional Independence (MCI)**: For each link $X_{t-\tau}^{(i)} \to X_t^{(j)}$ retained from phase 1, the MCI test (hypothesis in Eq. 1 ) further evaluates the dependence of $X_t^{(j)}$ on $X_{t-\tau}^{(i)}$, conditioned on the candidate parents of both nodes $X_t^{(j)}$ and $X_{t-\tau}^{(i)}$ (excluding the nodes themselves):

$$X_t^{(j)} \not\perp X_{t-\tau}^{(i)} \mid \mathrm{Pa}\left(X_t^{(j)}\right) \setminus \left\{X_{t-\tau}^{(i)}\right\}, \mathrm{Pa}\left(X_{t-\tau}^{(i)}\right). \quad (1)$$

If after conditioning on the parent sets, the dependence is no longer significant, the link $X_{t-\tau}^{(i)} \to X_t^{(j)}$ is pruned. Conditioning on parents of both nodes increases effect size under autocorrelation and rules out indirect/spurious links.

PCMCI thus estimates a time-lagged causal graph; our **causal feature set** is defined as the putative *direct lagged parents of the target* on this graph under PCMCI's assumption. For this study, we identify the lagged parents of electrical load as feature subset with `Tigramite`'s PCMCI implementation.

### B. Non-Causal Feature Selection Baseline

As a non-causal baseline, we (i) rank predictors by *Mutual Information* (MI) with the target, and (ii) remove near-duplicates via a correlation screen. MI is selected here as it captures both linear and nonlinear association. We keep important predictors whose MI exceeds an empirical threshold $MI > 0.025$. To limit redundancy, we then greedily remove predictors whose absolute Pearson correlation with any already-kept feature exceeds $|\rho| = 0.8$, a threshold

979-8-3315-8138-1

TABLE I
FEATURE SUBSETS.

| Feature | $F_0$ | $F_1$ | $F_2$ | $F_3$ |
|---|---|---|---|---|
| Load history | ✓ | ✓ | ✓ | ✓ |
| Datetime | ✓ | ✓ | ✓ | ✓ |
| Premise Count | ✓ | ✓ | ✓ | ✓ |
| Total Cloud Cover (`tcc`) | × | ✓ | × | × |
| Total Column Water (`tcw`) | × | ✓ | ✓ | × |
| Earth Surface Temperature (`skt`) | × | ✓ | × | ✓ |
| Net Long-Wave Radiation Flux (`avg-snlwrf`) | × | ✓ | ✓ | × |
| Net Short-Wave Radiation Flux (`avg-snswrf`) | × | ✓ | ✓ | × |
| 2-Meter Air Temperature (`t2m`) | × | ✓ | ✓ | ✓ |
| 2-Meter Dewpoint Temperature `d2m` | × | ✓ | × | × |
| Total Precipitation (`tp`) | × | ✓ | ✓ | ✓ |

chosen according to the *Variance Inflation Factor* (VIF) multicollinearity guideline $VIF = \frac{1}{1-\rho^2} \approx 2 < 5$.

### C. Evaluation Pipeline

We evaluate if feature selection improves city-level STLF using different models, each under controlled feature sets.

*1) Data Split:* We use sliding-window *time-series cross-validation* [23]: each test block is preceded by a fixed training window and an inner validation slice for early stopping; performance is averaged across folds. Scalers are fitted on training data only and applied to validation and test. (In our experiments, each sliding window spans 2 years, and we have 6 folds in total.)

*2) FS and STLF Setup:* For each city $c$, we train an individual model to predict that city's load using a 1-week history and a 24-hour horizon. We benchmark five models, i.e., SARIMAX, RF, GRU, TCN, and PatchTST, covering different types of well-established models. Together they span the design space of models that are most commonly used in short-term load forecasting: classical (*statistical*, or *classical ML*), and deep (*recurrent*, *convolutional*, *attention-based*) models. We compare four feature regimes: $\mathbf{F_0}$ **(Electricity-only):** Calendar/time features and load history from IESO; $\mathbf{F_1}$ **(All):** $F_0$ plus all ERA5 weather features (no selection); $\mathbf{F_2}$ **(Non-causal):** $F_0$ plus subset of ERA5 weather variables selected by the non-causal filter; $\mathbf{F_3}$ **(Causal):** $F_0$ plus subset of ERA5 weather variables selected by PCMCI algorithm, note that here we interpret $F_3$ as putative direct lagged parents of electricity demand under PCMCI's assumptions (*Causal Markov*, *Faithfulness*, and *no unobserved confounding*), not proven causal effects.

*3) Diagnostics:* We report: **Accuracy Across Feature Regimes:** Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), averaged over rolling-origin folds, reported across model classes and feature regimes; **Out-of-Distribution (OOD) Weather Robustness:** Performance is evaluated on extreme quantiles (e.g., $5^{th}$/$95^{th}$) of key weather drivers (temperature, precipitation). Models trained with causality-aligned features are expected to be more robust under such scenarios.

## V. EXPERIMENTS

### A. Datasets

We pair administrative-level electricity consumption with physically consistent hourly weather reanalysis for Ontario.

*1) IESO Ontario Hourly Electricity Consumption:* We use the Independent Electricity System Operator (IESO)'s public report, "Hourly Consumption by Forward Sortation Area (FSA)". We use records from January 2018 to March 2024, and have the following fields: `FSA` (first 3 characters of a Canadian postal code), `timestamp` (hour-ending), `consumer type` (*Residential* or *Small General Service*), `premise count` (total number of end users recorded in the past hour period), `total consumption` (kWh). In this work, we restrict to the *Residential* sector, as residential load reflects human activities and is most directly impacted by the weather conditions. Data entries are aggregated by FSA to the municipal levels to form city-scale residential load series. We choose regions of varying sizes and population levels, including: Toronto, Peel Region (Mississauga, Brampton), Hamilton, Brantford, Waterloo Region (Cambridge, Kitchener, Waterloo), London, Oshawa, Kingston, Ottawa. Geographical information of each city/region (longitudes, latitudes) is subsequently obtained to facilitate ERA5 data extraction.

*2) ERA5 Reanalysis:* ERA5 [3] is the $5^{th}$-generation global weather reanalysis data offered by Copernicus Climate Change Service (C3S), providing hourly estimates of a lot of atmospheric, land, and ocean variables. In this work, we use *ERA5 Single Level* [24] and *ERA5-Land* [25] datasets. The following covariates are extracted as potential causal drivers for analysis: total cloud cover (`tcc`), total column water (`tcw`), Earth surface temperature (`skt`), net terrestrial radiation flux (`avg-snlwrf`), net solar radiation flux (`avg-snswrf`), 2-meter air temperature (`t2m`), 2-meter dewpoint temperature (`d2m`), and total precipitation (`tp`). We use ERA5 as retrospective exogenous inputs to conduct feature selection.

### B. Feature Selection Results

Following Section III-B, we apply PCMCI to ERA5 reanalysis data and obtain the feature subsets reported in Table I. For subsequent experiments, we keep these subsets fixed and use them to define inputs available to each forecastor. This serves as proof of concept and can help showcase the benefits of causal feature selection. In an operational setting, the causal subset is first learned offline, and should be updated on a schedule (e.g., monthly) or when performance monitoring indicates degradation (hence, potential distribution drift).

The datetime features shared across all subsets are hour-of-day, day-of-week, month-of-year, day-of-year, and an Ontario holiday flag. In our implementation, cyclic fields (hour, day, month) are encoded with sine/cosine pairs to avoid artificial discontinuities. We also include `premise count` from IESO as a slow-moving proxy for the number of customers in each region.

**Compare** $F_2$ **and** $F_3$: Both $F_2$ and $F_3$ select `t2m` and `tp`. This is anticipated given the well-documented link, temperature→load, and also the role of rain or snow as proxies for weather regimes that alter occupancy and HVAC usage. Their selections diverge on radiation and moisture variables: $F_3$ drops `tcc` and `tcw` and retains `skt`; while $F_2$ keeps `tcw`, `avg-snlwrf`, and `avg-snswrf`. As a

979-8-3315-8138-1

TABLE II
CITY-WISE MAE (MWH) AND MAPE (%) FOR GRU, TCN, AND PATCHTST UNDER FOUR FEATURE REGIMES.
"TOP MAE/MAPE" COUNTS THE NUMBER OF CITIES WHERE A REGIME ATTAINS THE BEST SCORE FOR THE GIVEN MODEL.

| | | Toronto | | Peel | | Hamilton | | Brantford | | Waterloo | | London | | Oshawa | | Kingston | | Ottawa | | Count | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Feature Set** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **MAE** | **MAPE** | **Top MAE** | **Top MAPE** |
| SARIMAX | $F_0$ | 99.95 | 17.32 | 93.62 | 28.52 | 28.26 | 20.76 | 7.02 | 18.19 | 33.90 | 19.48 | 23.85 | 19.29 | 10.23 | 17.92 | 9.04 | 17.20 | **57.09** | 18.40 | 1 | 0 |
| | $F_1$ | **99.93** | **17.30** | 93.67 | 28.55 | 28.23 | 20.71 | **7.01** | 18.16 | 29.78 | **19.20** | 23.83 | 19.28 | 10.22 | 17.91 | 9.06 | 17.22 | 57.14 | **18.39** | 2 | 3 |
| | $F_2$ | 99.94 | 17.31 | 93.70 | 28.56 | 28.24 | 20.72 | 7.02 | 18.17 | 34.04 | 19.53 | 23.83 | 19.28 | 10.22 | 17.44 | 9.05 | 17.21 | 57.11 | **18.39** | 0 | 1 |
| | $F_3$ | 99.94 | **17.30** | **93.27** | **28.37** | **28.22** | **20.70** | **7.01** | **18.15** | **29.38** | 19.33 | **23.82** | **19.25** | **10.21** | **17.43** | **9.04** | **17.20** | 57.15 | **18.39** | **7** | **7** |
| RF | $F_0$ | 60.61 | 11.60 | 45.73 | 14.62 | 11.23 | 7.18 | 3.37 | 7.73 | 10.96 | 6.45 | **8.11** | **6.00** | 5.77 | 8.44 | 4.44 | 8.62 | 30.32 | 8.80 | 1 | 1 |
| | $F_1$ | 59.10 | 11.30 | 41.93 | 13.23 | 11.17 | 7.12 | 3.14 | 7.43 | 10.47 | 6.27 | 8.59 | 6.22 | 5.87 | 8.61 | 4.40 | **8.41** | 30.77 | 9.01 | 0 | 1 |
| | $F_2$ | 59.40 | 11.38 | 40.15 | 10.36 | **10.87** | **6.99** | **3.01** | **7.13** | 10.49 | 6.31 | 8.34 | 6.07 | **5.69** | **8.37** | 4.42 | 8.49 | 29.85 | 8.76 | 3 | 3 |
| | $F_3$ | **58.05** | **11.08** | **39.32** | **10.14** | 10.89 | **6.99** | 3.03 | 7.18 | **10.36** | **6.22** | 8.33 | 6.05 | 5.72 | 8.41 | **4.39** | 8.43 | **29.80** | **8.74** | **7** | **5** |
| GRU | $F_0$ | 29.21 | 5.06 | 16.81 | 5.28 | 7.55 | 5.15 | 2.18 | 5.44 | 7.71 | 4.97 | 7.79 | 6.19 | 4.21 | 6.80 | 3.24 | 6.78 | 22.33 | 6.91 | 0 | 0 |
| | $F_1$ | 29.90 | 5.12 | 17.12 | 5.45 | 8.12 | 5.29 | 2.22 | 5.37 | 8.17 | 5.17 | 7.80 | 6.16 | 4.19 | 6.63 | **3.14** | **6.53** | 23.98 | 7.11 | 1 | 1 |
| | $F_2$ | 29.43 | 5.03 | 16.75 | 5.16 | **7.50** | **5.05** | **2.16** | **5.26** | 7.76 | 4.93 | 7.11 | 5.85 | 3.91 | 6.35 | 3.25 | 6.74 | 22.76 | 6.80 | 2 | 2 |
| | $F_3$ | **29.17** | **5.01** | **15.82** | **4.89** | **7.50** | **5.05** | 2.20 | 5.32 | **7.51** | **4.76** | **6.84** | **5.34** | **3.84** | **6.16** | 3.23 | 6.64 | **21.27** | **6.53** | **7** | **7** |
| TCN | $F_0$ | 27.06 | 4.75 | 17.60 | 5.56 | 8.25 | 5.59 | 2.19 | 5.51 | 8.06 | 5.27 | 5.97 | 4.65 | 3.64 | 6.11 | 3.81 | 7.43 | 24.00 | 7.31 | 0 | 0 |
| | $F_1$ | 26.88 | 4.64 | **17.08** | **5.27** | 8.63 | 5.74 | 2.28 | 5.67 | 8.49 | 5.40 | 5.94 | 4.51 | 3.68 | 6.07 | 3.61 | 7.19 | 23.60 | 7.27 | 1 | 1 |
| | $F_2$ | **26.40** | **4.55** | 17.70 | 5.55 | 8.74 | 6.00 | **2.00** | **4.96** | **7.58** | **4.89** | 5.98 | 4.52 | 3.53 | 5.86 | 4.12 | 8.17 | **22.20** | **6.79** | **4** | **4** |
| | $F_3$ | 26.53 | 4.60 | 17.11 | 5.34 | **8.47** | **5.72** | 2.13 | 5.20 | 7.82 | 5.18 | **5.83** | **4.49** | **3.25** | **5.47** | **3.55** | **7.04** | 25.10 | 7.70 | **4** | **4** |
| PatchTST | $F_0$ | 26.32 | 4.55 | 14.20 | 4.56 | 6.95 | 4.85 | 1.83 | 4.75 | 7.48 | 4.88 | **5.45** | 4.35 | 2.67 | 4.64 | 2.49 | 5.07 | 15.50 | 5.05 | 1 | 0 |
| | $F_1$ | 25.40 | 4.44 | 15.73 | 5.08 | 7.63 | 5.26 | **1.81** | **4.63** | **7.03** | **4.54** | 5.90 | 4.69 | 2.74 | 4.69 | 2.32 | 4.70 | 15.40 | 4.99 | 2 | 2 |
| | $F_2$ | 24.53 | 4.33 | **14.15** | **4.53** | 7.06 | 4.89 | 1.88 | 4.84 | 7.31 | 4.77 | 5.72 | 4.53 | 2.72 | 4.56 | 2.54 | 5.21 | **14.90** | **4.91** | 2 | 2 |
| | $F_3$ | **24.37** | **4.28** | 14.78 | 4.58 | **6.91** | **4.70** | 1.85 | 4.76 | 7.18 | 4.69 | **5.45** | **4.26** | **2.59** | **4.38** | **2.31** | **4.62** | 15.40 | 5.06 | **5** | **5** |


Fig. 2. Toronto's OOD Weather in Test Year: Heat, Cold Wave, Heavy Precipitation

causal discovery method, PCMCI identifies variables that potentially have a more direct effect on load and prunes features whose influence is indirect and mediated (e.g., cloud→radiation→temperature→load, or moisture→thermal comfort→temperature proxies→load). In this case study, radiation and column water lose significance once temperature variables are identified as the putative direct driver of load. In short, $F_2$ reflects a more correlational association: it keeps radiation and column water, as they explain a part of the variance of the prediction target. $F_3$ reflects structural parsimony: it favors more direct thermal drivers and discards variables whose effects are indirect or redundant after conditioning.

### C. City-Level Forecasting Results

We first train and evaluate all models with ERA5 reanalysis as a weather proxy. Though assuming perfectly accurate weather input is an overly ideal assumption in real-world settings, this serves as a proof-of-concept experiment.

Table II reports MAE (MWh) and MAPE (%) for each city under the four feature regimes ($F_0$–$F_3$). In general, PatchTST outperforms the other models across all cities with the lowest MAE/MAPE scores. For every model, $F_3$ attains the most city-wise best scores (by the count of lowest MAE/MAPE), showing systematic gains from pruning indirect meteorological covariates. The results also confirm that using all features is not always the best: $F_1$ overall underperforms $F_2$ and $F_3$, suggesting that naively adding all weather variables may introduce spurious correlations and harm performance. Across architectures, selecting putative *direct* drivers ($F_3$) generally yields modest but reliable accuracy improvements over $F_0$/$F_1$, and often outperforms a non-causal filter ($F_2$). Overall, $F_3$ tends to win across all benchmark models, suggesting that causal FS is a model-agnostic way to integrate weather while controlling feature-set complexity.

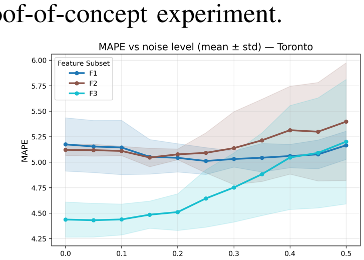


Fig. 3. Toronto Forecasting: Noise Sensitivity with PatchTST's Different Feature Regimes.

**Noise Sensitivity.** In actual forecasting deployment, the weather inputs available are usually *forecast* products, thus inevitably contain more errors than reanalysis data. To assess robustness to imperfect weather inputs, we perform a controlled noise-injection sensitivity analysis. Specifically, we keep the trained forecaster fixed and perturb only the ERA5 covariates at test time. For each weather variable $x$, we apply the multiplicative perturbation $\tilde{x} = x\cdot(1+\epsilon), \quad \epsilon \sim \mathcal{U}(-\alpha, \alpha)$, where $\alpha$ controls the relative noise magnitude. We sweep $\alpha \in \{0, 0.05, 0.10, 0.15, \ldots, 0.40, 0.45, 0.50\}$, corresponding to perturbations of up to $\pm 50\%$ of each variable's magnitude. Focusing on feature subsets $F_1$, $F_2$, and $F_3$ (which include weather inputs), we train PatchTST on Toronto using unperturbed ERA5 over 2018-01-01 to 2023-03-10, and evaluate on the held-out test year (2023-03-11 to 2024-03-10) using noise-contaminated ERA5, with the same prediction horizons as in Section V-C. We repeat the perturbation with 5 random seeds and plot MAPEs (average and standard deviation across seeds) in Fig. 3. While $F_3$ exhibits a slightly steeper degradation in MAPE than $F_1$ and $F_2$ (this is exactly consistent with the fact that $F_3$ contains the most influential causal weather drivers), $F_3$ still achieves the lowest MAPE across all noise levels, indicating improved noise robustness under the causally selected feature subset.

### D. Out-of-Distribution (OOD) Weather Inference Results

We define OOD weather relative to the history for each city: compute the $5^{th}$/$95^{th}$ percentiles of `t2m` and `tp` using

979-8-3315-8138-1

TABLE III

DEEP FORECASTORS ON OOD-WEATHER. METRICS: MAE (MWH) AND MAPE (%). BOLD: THE BEST FOR EACH CITY ACROSS ALL CONFIGURATIONS. UNDERLINE: $2^{nd}$ BEST. COMPUTE RELATIVE ERROR REDUCTION (%) FROM $2^{nd}$ BEST TO THE BEST.

| Model | Feature Set | Toronto MAE | Toronto MAPE | Ottawa MAE | Ottawa MAPE |
|---|---|---|---|---|---|
| GRU | $F_0$ | 45.62 | 6.43 | 38.25 | 9.70 |
| | $F_1$ | 48.35 | 6.87 | 27.23 | 7.40 |
| | $F_2$ | 45.46 | 6.48 | 34.13 | 8.93 |
| | $F_3$ | 47.65 | 6.66 | 38.89 | 9.74 |
| TCN | $F_0$ | 44.73 | 6.24 | 29.19 | 7.31 |
| | $F_1$ | 45.15 | 6.30 | 26.48 | 6.83 |
| | $F_2$ | 43.53 | <u>6.10</u> | 26.10 | 6.89 |
| | $F_3$ | 44.88 | 6.31 | 27.12 | 7.05 |
| **PatchTST** | $F_0$ | <u>42.13</u> | **5.70** | 26.13 | 6.73 |
| | $F_1$ | 47.79 | 6.84 | <u>24.80</u> | <u>6.60</u> |
| | $F_2$ | 47.81 | 6.64 | 26.41 | 7.13 |
| | $F_3$ | **40.13** | **5.70** | **23.69** | **6.24** |
| *Err. Reduction (%)* | | *4.75* | *6.56* | *4.48* | *5.45* |

new training window, 2018-01-01 to 2023-03-10. Any 24h window in the held-out test year (2023-03-11 to 2024-03-10) is flagged OOD if over 50% of its 24 hours fall outside those thresholds, with a minimum 24 h between windows. We report OOD windows for Toronto and Ottawa (E.g., Toronto OODs in Fig. 2) in the test year with 3 deep forecastors, GRU, TCN and PatchTST, each retrained from scratch on the full training window (same hyperparameters as in Section V-C). We evaluate these models on OOD windows across feature regimes and report average MAE and MAPE across all OOD windows (Table III). Overall, the best configuration in both cities is PatchTST with $F_3$ (PCMCI), beating all other configurations. This suggests that pruning indirect or redundant weather covariates via causal selection yields more robust forecast under extreme weather conditions, particularly for attention-based models that benefit from reduced input redundancy. On the other hand, GRU and TCN show mixed behavior, occasionally favoring $F_0$/$F_2$, suggesting architecture-specific interactions with feature sparsity and redundancy that merit further study. Still, across cities and different weather OODs, PatchTST with feature set $F_3$ is consistently the strongest, supporting our claim that causal FS improves robustness under distribution shift while maintaining accuracy.

## VI. CONCLUSION

This case study evaluates if causal feature selection enhances weather-informed STLF. Across 9 Southern Ontario regions and multiple load forecasting models, selecting features via PCMCI generally yields a more compact input feature set, improves accuracy, and can perform more robustly under imperfect weather observations and under extreme weather events. Overall, PCMCI causal discovery method can be a model-agnostic, lightweight module that improves load forecasting performance while strengthening generalization under distribution shifts. Future work will expand to multi-city multivariate forecasting, causal transfers across regions, and introducing contemporaneous effects via PCMCI+ to further stress-test causal feature selection in operational settings.

## REFERENCES


[1] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[2] M. G. Pinheiro, S. C. Madeira, and A. P. Francisco, "Short-term electricity load forecasting—a systematic approach from system level to secondary substations," *Applied Energy*, vol. 332, p. 120493, 2023.

[3] C. Soci, H. Hersbach, A. Simmons, P. Poli, B. Bell, P. Berrisford, A. Horányi, J. Muñoz-Sabater, J. Nicolas, R. Radu *et al.*, "The era5 global reanalysis from 1940 to 2022," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 764, pp. 4014–4048, 2024.

[4] F. Sheng and L. Jia, "Short-term load forecasting based on sarimax-lstm," in *2020 5th International Conference on Power and Renewable Energy (ICPRE)*. IEEE, 2020, pp. 90–94.

[5] G. Dudek, "A comprehensive study of random forest for short-term load forecasting," *Energies*, vol. 15, no. 20, p. 7547, 2022.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[7] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[8] Y. Nie, "A time series is worth 64words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[10] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[11] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 international conference on engineering & MIS (ICEMIS)*. IEEE, 2016, pp. 1–5.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[13] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, and R. García-Herrera, "Feature selection in machine learning prediction systems for renewable energy applications," *Renewable and Sustainable Energy Reviews*, vol. 90, pp. 728–741, 2018.

[14] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation." *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[15] D. M. Hausman and J. Woodward, "Independence, invariance and the causal markov condition," *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

[16] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[17] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[18] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS*, vol. 2, 2003, pp. 376–81.

[19] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on time series using restricted structural equation models," *Advances in neural information processing systems*, vol. 26, 2013.

[20] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science advances*, vol. 5, no. 11, p. eaau4996, 2019.

[21] J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Conference on uncertainty in artificial intelligence*. Pmlr, 2020, pp. 1388–1397.

[22] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

[23] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.

[24] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum *et al.*, "Era5 hourly data on single levels from 1940 to present," *Copernicus climate change service (c3s) climate data store (cds)*, vol. 10, 2023.

[25] J. Muñoz Sabater *et al.*, "Era5-land hourly data from 1981 to present," *Copernicus climate change service (C3S) climate data store (CDS)*, vol. 10, no. 10.24381, 2019.


979-8-3315-8138-1