Drawing Direct and Indirect Causal Relationships Between Climate Variables

Statistical Machine Learning, Fall 2025

Kelsey Hawkins

Dr. Wissam Ahmed

May 16th, 2025

# Abstract

Causal inference must be approached specifically, with a robust background and framework about the topic. Bringing causal inference and machine learning into water usage analysis is a task that has not been approached in this specific way. Utilizing double machine learning (DML), to bridge the gap between statistics and machine learning, will allow the analysis and determine causal relationships between climate variables and water usage in a residential setting. A key concept in causal inference is that correlation does not imply causation. Finding correlations between variables is the beginning step to causal analysis, to determine previous knowledge to test as something to accept or reject.

The models utilized are causal forests, with XGBoost regressors as the building block of each causal tree to build a robust non-linear causal analysis framework to answer our question. Causal forests are an ensemble model, building off of the random forest ensemble tree model. The causal forest is an ensemble of causal trees, building off of a decision tree. The tree is instead split on the biggest difference between treatment effects in the node of comparison. This allows us to take the average treatment effect (ATE) over all of the leaves in the tree to find the average treatment effect for each tree. XGBoost regressor is utilized within each causal tree to account for confounders by minimizing the residuals of the model, given each split from the tree nodes. Taking this double machine learning approach makes our analysis robust and unbiased, resulting in an overall average treatment effect (ATE) for each treatment variable in relation to water usage. Since a direct and indirect causal effect will be tested, many causal forests will be utilized to test for each direct and indirect causal relationship. The relationships will be visualized and denoted utilizing a Directed Acyclic Graph (DAG) to visualize the causal pathways between variables and water usage.

TODO - add primary results

TODO - significance/implications of findings

# Introduction

Defining a causal relationship is a difficult task, especially when it comes to finding a causal relationship between climate factors and water usage. Utilizing a simple statistical approach can only yield at most, a correlation analysis between variables which does not imply causality. The traditional statistical approach does not offer a robust approach to assessing causal relationships between the treatment variable and predictors, thus the necessity for double machine learning and causal forests.

The objective of this study is to determine the causal relationship between five climate factors and residential water usage. These factors include: relative humidity, wind speed, relative air temperature, precipitation, and evapotranspiration. By utilizing an approach that combined double machine learning and causal inference, a causal inference result can be analyzed and concluded. This approach was chosen to combine the strength of statistics, machine learning, and causal inference together. Since the output of these models provide treatment effects, we can analyze which climate factors have the most effect on water usage, as treatment changes.

Determining the most influential climate factor on water usage is important for water districts to gear their analyses and actions towards water conservation. Finding the variables that influence water usage the most will help find causal pathways that will open the doors to future research and current actions that can be taken to promote water saving based on these results. This fits in the broader field of machine learning as incorporating machine learning into causal inference, a natively statistical topic, will help bring causal inference studies into use more in many fields.

The next section provides a literature review, followed by the methodology, results, and conclusions. Each section will go in depth about each topic discussed and provide detailed analysis and discussion.

# Literature Review

Recent advances in causal inference have introduced powerful tools for estimating heterogeneous treatment effects (HTEs) in observational data, particularly through machine learning frameworks. One prominent development is the causal forest (Wager & Athey, 2018), an extension of random forests designed for causal effect estimation. Unlike traditional predictive models, causal forests focus on estimating conditional average treatment effects (CATEs), leveraging subsampling and honest splitting to reduce overfitting and bias. These models have been instrumental in moving beyond average treatment effect (ATE) estimates to more granular insights into how treatment effects vary across subpopulations.

Building on this, Double Machine Learning (DML) frameworks (Chernozhukov et al., 2018) have gained traction for addressing confounding biases in high-dimensional settings. DML methods decouple the estimation of nuisance parameters (e.g., propensity scores, outcome models) from the final causal effect estimation using orthogonalization. This two-step process allows for valid inference even when complex machine learning models are used for nuisance estimation. When combined with causal forests, DML can further enhance the robustness of treatment effect estimates by explicitly modeling both outcome and treatment assignment processes using flexible learners.

Recent literature has begun to explore hybrid frameworks where gradient boosting methods like XGBoost (Chen & Guestrin, 2016) serve as the base learner within each tree of a causal forest or DML algorithm. XGBoost's regularization, tree pruning, and scalability make it a particularly effective regressor in high-dimensional settings and non-linear data structures. Several empirical studies have demonstrated that using boosted trees within causal forests improves estimation precision and reduces variance in effect estimates (Künzel et al., 2019). Moreover, the combination of XGBoost with DML techniques enables researchers to model complex interactions between covariates and treatments while preserving the interpretability of estimated treatment heterogeneity.

Applications of these methods span a wide range of domains, including economics, public policy, and environmental science. For instance, Künzel et al. (2019) applied causal forests to personalize medical treatments, while Nie and Wager (2021) proposed a generalized framework for estimating heterogeneous effects using orthogonalized loss functions. These approaches have proven particularly valuable in settings with limited experimental data, where understanding localized treatment effects is crucial for policy decisions. The integration of tree-based machine learning algorithms into causal inference workflows represents a promising direction for producing actionable, data-driven insights in observational studies.

# Methodology

## Data Description

The dataset was provided by Irvine Ranch Water District (IRWD), originally utilized in previous research. This research builds off of previous research completed as we determine indirect causal pathways between variables instead of simple direct causal relationships, as well as expanding data utilized in the models to account for more variance in causal impacts. The dataset has six years of data consisting of customer billing data and climate data. A monthly scale is utilized, each row representing a customer's residential data for a given month of a given year. There are over 100,000 residential households present in the data. The data preprocessing was robust and occurred over one year before this research was completed. Originally, the dataset was extremely messy with variables not needed for research purposes, and they were removed. The data was on an uneven scale as not every billing cycle is the same number of days. Each numerical time series datapoint was normalized by the number of billing days in the cycle that was present for each customer. This allows for data to be analyzed on the same time scale. All null and missing values were filled in with 0's as each row is precious data that can not be dropped. There are no sampling methods or extra features created. The dataset is over 2.5 GB in size, with over six million rows of data.

## Model Selection / Development

Double Machine Learning (DML) is a framework designed to estimate causal effects in the presence of high-dimensional confounders by separating the estimation of nuisance parameters (e.g., the treatment and outcome models) from the final causal parameter of interest. When integrated with causal forests, this approach allows for the estimation of heterogeneous treatment effects (HTEs) by leveraging ensemble-based non-parametric models that can flexibly model non-linear relationships and complex interactions. In this setup, XGBoost serves as the base regressor within each tree of the causal forest, improving prediction accuracy through gradient boosting and regularization. The process begins by using XGBoost to estimate the conditional expectation of the outcome and the treatment assignment (propensity score). These nuisance estimates are then orthogonalized—removing their influence from the treatment effect estimation—to reduce bias and ensure valid inference. The causal forest subsequently estimates treatment effects at the individual or subgroup level by partitioning the feature space and averaging effects across similar observations, where each partition relies on accurate base learners. This combined framework of DML, causal forests, and XGBoost provides a robust, interpretable, and scalable method for uncovering nuanced causal relationships in observational data.

## Training and Validation Strategy

Since this causal machine learning model requires all of the data, it was not split into training and testing sets, as a normal machine learning model is utilized. The metrics utilized to validate this model is estimating the causal effects using Average Treatment Effects (ATE). Since the true causal effect is generally unknown in observational data, validation becomes impossible or less straightforward than predictive models that are commonly utilized. Causal forests inherently use sample splitting which is a form of cross-fitting to maintain honesty within trees. They use half of the data to build the tree structure and the other half to estimate effects. It reduces overfitting and gives valid confidence intervals for treatment effects. The double machine learning approach uses cross-fitting where parameters are estimated in one fold and used in another to compute treatment effects. We can utilize RMSE or MSE for the outcome regression for the XGBoost regressor however it gets complicated due to the inside nesting of the regressor model within the causal tree. They do not actually determine if the model is doing well as causal inference does not have a true value to compare to. Visual diagnostics are much more reliable for example histograms of ATE estimates, calibration plots, treatment effects vs covariate plots, confidence intervals, or propensity score diagnostics all can be utilized here.

## Implementation Details

The libraries utilized are the custom library of the Causal forest, XGBoost regressor from Sklearn, and computational resources from Chapman University's GPU servers.

## Experimental Setup

In order to assess causal relationships between climate variables and water usage, we conduct many experiments using the causal forest DML estimator from the econml library in Python. This model combines double Machine Learning (DML) framework with nonparametric causal forests, for the estimation of heterogeneous treatment effects (HTEs) in the observational data. The climate variables include : relative humidity, wind speed, air temperature, precipitation, and evapotranspiration. Each climate variable was treated as a separate continuous treatment variable within each model to isolate direct effects. Indirect relationships are later explored through sequential modeling.

For the flexibility and robustness of the model, XGBoost regressors were used for the outcome model and treatment (propensity) model. Cross-fitting and sample splitting were implemented within the causal forest estimator to control for overfitting and to produce unbiased treatment effect estimates. For each model we compute the Conditional Average Treatment Effects (CATEs) and Average Treatment Effects (ATE). Since there is no direct ground truth for causal effects in observational data, the model's internal diagnostics - such as histogram plots of CATE distributions and confidence intervals - provided validation and interpretability.

All models and code are implemented in Python using the econml, xgboost, and scikit-learn libraries and computations were carried out on GPU-backed servers provided by

Chapman, due to the size and model complexity. This experimental setup enables efficient estimation of causal effects across the large dataset and desired treatment variables.

# Results - TODO

(I will finish by part C)

# Discussion - TODO

(I will finish by part C)

# Conclusion - TODO

(I will finish by part C)

# References - TODO

(I will finish by part C)


- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. The Econometrics Journal, 21(1), C1–C68.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). *Metalearners for estimating heterogeneous treatment effects using machine learning*. Proceedings of the National Academy of Sciences, 116(10), 4156–4165.
- Nie, X., & Wager, S. (2021). *Quasi-oracle estimation of heterogeneous treatment effects*. Biometrika, 108(2), 299–319.
- Wager, S., & Athey, S. (2018). *Estimation and inference of heterogeneous treatment effects using random forests*. Journal of the American Statistical Association, 113(523), 1228–1242.