The Pennsylvania State University
The Graduate School

TOWARDS A GEO-AGNOSTIC, SOURCE-AGNOSTIC MODELING OF CLIMATE INFLUENCES ON RENEWABLE POWER PLANT-LEVEL GENERATION

A Thesis in
Energy and Mineral Engineering
by
Vijay Bhaskar Chiluveru

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2024

The thesis of Vijay Bhaskar Chiluveru was reviewed and approved by the following:

Renee Obringer
Assistant Professor of Energy and Mineral Engineering
Thesis Advisor

Russell R Barton
Distinguished Professor of Supply Chain and
Information Systems in Smeal College of Business

Mort D Webster
Professor of Energy and Mineral Engineering

Jeremy Gernand
Associate Professor of Energy and Mineral Engineering
Chair of the Graduate Program

Abstract

Energy infrastructure is critical to modern society. However, the ongoing climate crisis is already impacting existing energy infrastructure through extreme weather events which are increasingly frequent and intense. In fact, these climate-induced impacts may create roadblocks for the energy transition, particularly if the climate impacts on low carbon and renewable energy technologies are not well-understood. Here, I propose a data-driven methodology to model these complex interactions defining the renewables-climate-risk nexus over large spatiotemporal scales. In particular, this study leverages an open-source dataset containing hydro, wind and solar energy systems across the United States. Using tree-based ensemble learning techniques, it is shown that we can model the non-linear effects of climate variables on these renewable systems. This study further demonstrates the potential of training Random Forests to produce geo-agnostic, source-agnostic models which are aimed to have consistent and comparable performance with respect to source-specific modeling. This study and research work is aimed at envisioning a future, in line with the trends of rising renewables in the mix, with the explicit research need to look at common data pipeling and modeling frameworks in developing data-driven models which can be geo-agnostic as well as source-agnostic to aid with the long-term planning and operations of energy systems in a future actively impacted by climate change.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Obringer, for believing in me and granting this project to work on. Her constant guidance and feedback in the progress of my research work has shaped me to be more proactive towards data analysis and well versed in applying statistical learning techniques. With her, I had the opportunity to present my research at multiple avenues where I got to interact with many experts of the field and expand my research horizon. I have a deep sense of gratitude to her for everything she has done for me and also, for everything she has encouraged me to pursue here at Penn State.

I wish to extend special thanks to Dr. Barton and Dr. Webster for seeing interest and potential in my research work and providing me their valuable feedback, helping me mould this project into an effective piece of work. Their support, along with the directions from my advisor, has allowed me to present this work in its fullest form.

My family also deserves special appreciation for their unwavering belief in me and extending their constant support all the way from India. Finally, I take this opportunity to give a huge shout-out to my friends everywhere for cheering me up on my lowest days and celebrating my wins. I extend my heartfelt thanks for their warmth as I toil away in this cold climate.

Chapter 1
Introduction

We are already witnessing significant impacts of climate change in our collective daily lives. This is especially pronounced in terms of ongoing paradigm shifts in energy demand, consumption and generation [1]. This crisis has affected energy infrastructure across the world through intense hot and cold temperatures as well acute precipitation-related events like droughts and floods. Increasing frequency as well as magnitudes of such extreme weather events and multi-hazard phenomena, majorly attributable to anthropogenic greenhouse gas emissions, are anticipated to rise in frequency over the coming decades [2–4]. This is resulting in a worldwide trend of global electricity generation and transmission systems experiencing stress for increasing periods of operation [1, 5, 23]. Incidentally, our global energy needs are one of the largest contributors to these emissions, with almost a two-thirds share coming from only energy sector in 2019 [7]. Intergovernmental Panel on Climate Change (IPCC) studies have shown that this energy-climate-nexus feedback loop where energy system resiliency is limited by the manifestations of the same sector's contribution to climate change [8]. This has necessitated a rethinking of energy systems planning and operations to ensure climate-resilient and cost-effectively reliable modes of electric supply to our societies [1, 5, 54].

1.1 Climate Change Impacts on Energy Systems

Accelerating the pace of adopting renewable energy sources is part of this transformation efforts of energy sector [10]. However, these sources are generally more intermittent and highly contingent on ambient climatic conditions [11, 12], meaning the above mentioned climate-induced bottlenecks to energy systems vulnerabilities are only going to become more influential on society's mitigation and adaptation efforts. Thus, energy system modeling integrated with climate variability modeling is required to understand climate

impacts on power systems and enable robust short-term, medium-term and long-term operational planning [10, 13].

Currently, there is a substantial focus of research on energy-climate-nexus modeling. It has been observed that rapidly changing atmospheric, hydrological and other climatic conditions can affect every component of power system planning, from demand and supply to transmission and storage [14, 15, 23]. Major studies investigating the patterns of climate change present its varying effects in only two geographic regions, Europe and North America [14, 52–54, 59, 61]. The study projections provide an understanding of the effects of climate change based on changing wind patterns and speeds, land heating, precipitation and energy supply as well as consumption. The correlation of these factors with energy production, reliance on renewable energy resources, and the peak production and demand reveal that climate change will have a huge impact on the energy production and governments and energy/utility/distribution companies must include these effects while planning not only their short-term and medium-term but also long-term operations [55, 59–61, 63]. Instances of poor preparation have led to breakdown of systems during the 2021 Texas cold wave and studies display irregular weather in the North America region in the near future [52–54, 61]. To avoid this, energy systems must be prepared and leverage the projected increased sun days in Texas, precipitation in the north western regions, extreme winds and cold in different parts of the USA. Therefore, the increased effects of climate change in southern states like Texas and reduced effects in the northwest would result in increased and decreased loads in these regions [14].

Most of this research, however, has focused on impact quantification at the scale of one (or few) component(s) of the overall electric grid. Even within research focused on either the demand or supply aspect of energy, there are patterns varying considerably by geography and seasonality. For example, it has been observed that energy used in the USA for satisfying summer-time cooling needs is likely to increase due to climate change [17,19]. Broadly, rising surface and air temperatures is set to result in higher cooling demand and increased annual and peak electricity consumption [18]. Shifts in wind speed and relative humidity, in conjunction with increasing temperatures, have been found to affect building heating and cooling requirements [20,21] which give newfound variance to overall electricity use. These increases in demand connect back to the generation, which is also likely to shift under future climate change. For instance, drought-prone conditions will exacerbate hydropower production capabilities [63] while also sparking a cross-sectoral resource competition [22] but this can increase solar energy generation due to reduced cloud cover in the same region [12].

On the supply side, hydro, wind and solar power have all been found to have climate change induced variance in operational efficiencies due to precipitation, cloud cover, temperature and wind speed [12, 23–26, 52]. Most of the studies on generation capabilities have limited spatio-temporal scale and are more focused on regional impact assessments. The data available for analysis are from these limited spatial scales and developed economic regions and do not display the holistic aspects of changes across wide regions. Additionally, such similar data are only analysed for specific regions, specific times, and sometimes both for only one (or two) energy source. The energy generation from these resources are projected to be affected in a highly varied manner by climate change, in either positive or negative accounts depending on the study geography, time period and energy source [53, 57, 60, 63]. The increased heat days in the southern USA are expected to increase the production of solar energy but the increased winds and dust can ultimately reduce the generation in such photovoltaic processes [52–54]. Further, high winds are expected in the north and central regions of the USA which could substantially increase the wind energy production but the rate of projected winds present a cut-off scenario of turbine operations along with moving these windmills to west for much more efficient generation [55]. Such effects are found to decrease the production of energy by 5% in every cycle while simultaneously occuring in future periods of potentially increasing consumption and energy demand [55–57].

For hydro, review studies have observed that an overall global decrease in annual generation [1, 5, 28, 29] with strong regional and seasonal differences is forecasted for the near future. In contrast, studies on solar generation changes either indicate slightly positive or negatives shifts depending on the study's geography and time period [30–32]. Both increased and reduced wind power potential has been reported by studies [24, 33, 34], with little overall consensus on the macro-scale overall impact on the renewable subsector. Review papers on synthesizing generation-side climate impacts have shown that hydro, wind and solar (along with other renewables) have large gaps with respect to covering global geographies [1, 5, 35]. Few papers are comprehensive in terms of understanding climate change impacts on a multi-component-level. In addition to expanding the geographical scope of our models, it is critical that we increase the temporal resolution, zooming into subseasonal timescales, to fully understand how energy systems may be impacted by climate change [1, 34]. Due to the highly interconnected nature of renewable energy systems and their dependence on similar climate variables, it becomes important to develop robust and consistent models which can have good predictive power over a broad range of dimensions such as time, geography, and energy source type. This can

enable researchers and practitioners to build systems in place which can be useful in optimized operations and planning.

1.2 Research Gaps

Review papers on synthesizing generation-side climate impacts have shown that hydro, wind and solar (along with other renewables) have large gaps with respect to covering global geographies [1, 5, 35]. The absence of research about the effects of climate change on energy generation and consumption is observed majorly in the developing world [57]. For instance, the continent of Africa is projected to have many more heat days in the next thirty years and the opportunity of generating solar energy and satisfying the needs of energy consumption in different regions and times must be explored. Few papers are comprehensive in terms of understanding climate change impacts on a multi-component-level. The cascading effects also present a deeper understanding of how changing climate will generate difficulties in energy generation [58]. Any increased precipitation in regions of the USA has potential to boost the hydro energy generation but the increased requirement of water in summer and winter time by the regional population is expected to hit this energy production [58]. The increasing temperature of earth is also expected to heat the water, accelerate growth of flora in reservoirs, and therefore affect the generation through turbines leading to outages [61, 63].

In addition to expanding the geographical scope of our models, it is critical that we increase the temporal resolution, zooming into subseasonal timescales, to fully understand how energy systems may be impacted by climate change [1, 34]. Therefore, more research into understanding the effects of climate change on the medium-term and long-term generation capabilities renewable resources and portfolios is needed. Due to the highly interconnected nature of renewable energy systems and their dependence on similar climate variables, it becomes important to develop robust and consistent models which can have good predictive power over a broad range of dimensions such as time, geography, and energy source type. This can enable researchers and practitioners to not only augment existing energy systems modeling efforts but also to build novel predictive modeling systems. These data-driven models aimed at understanding how future climate change impacts can impact our renewables portfolio performance can be useful in optimizing long-term operations and planning. This studying will build towards more climate-resilient energy systems operations and planning to counteract uncertainties of climate change and ensure effective satisfaction of our collective energy needs [59, 60, 64].

1.3 Research Objectives

In this study, I aim to address the above discussed research gaps by leveraging a rich dataset encompassing the United States power plant portfolio to develop a *generalized* predictive model. This model can capture the non-linear renewables-climate interactions across time, geography and energy source. My data spans the whole contiguous geography of the USA from the time period of 2011-22 and it is hyper-local. That is, each datapoint corresponds to a particular power plant along with the corresponding climate variables data. Using this, I then develop tree-based ensemble learning models and test them for the potential to become a generalized model across the time period of 2011-22 and the predictive power to model Hydro, Wind and Solar generation across USA. This dataset and study aims to enable open-sourced, global collaboration into building sound models of renewable energy systems and contribute to the renewables-climate-risk modeling community. Summarizing the above, the primary objectives are:

1. Develop a rich and robust dataset containing downscaled climate variables and individual renewable power plant-level generation and capacity.

2. Validate the capacity-generation linear relationship benchmark and demonstrate increased performance of generation prediction models by capturing non-linear climate effects on generation.

3. Identify the modeling technique within (2) which is giving best performance in generation prediction.

4. Using the best technique in (3) to build a *generalized* model with CUF (Capacity Utilization Factor, discussed further in *methodology*) as the response variable and validate its performance using stratified test data.

5. Compare the generalized CUF model performance with source-specific (Hydro, Wind and Solar) models to illustrate potential of geo-agnostic, source-agnostic modeling.

1.4 Reader's Guide

The thesis is organized as follows. First, I discuss the data and methodology used in this study. Then, I delve into the results and discussion, focusing on a comparative analysis of performance of the models developed and show the best performer. The subsequent best model performance across time and energy-source stratified test data is to demonstrate the potential for realizing generalized predictive models. Finally, I discuss implications and opportunities for future work before concluding my research.

Chapter 2
Research Methodology

2.1 Data Description

In this study, I developed a novel dataset for the USA using two publicly available federal data sources for the time period of 2011-2022. The first is the electricity sector data and the second is the observational climate data.

The electricity sector data across the USA was obtained from the U.S Energy Information Administration (EIA). Two distinct subsets that were used for this are the EIA-860 [39] and EIA-923 [40] datasets. The former contains variables which define a particular power plant within the overall portfolio across the country. These include a uniquely assigned plant code, geographical markers of the plant location such as the state, county and the latitude/longitude of the plant's location, operational plant capacity (in MW) and the time since each power plant began its operations. Whereas, EIA-923 data contains the variables relevant to power plant generation like the source of energy and monthly generation (in MWh) along with the already mentioned plant code. Both these data sources are combined using the plant code variable as the primary key. Together, this data gives us the information to identify and locate any operational power plant across the United States, and analyze its monthly generation over the years.

The recorded historical climate data was collected from the North American Regional Reanalysis [41] (NARR). The daily values of the NARR data were aggregated to get monthly measures. This was done to maintain parity and temporal consistency with the energy sector data which has monthly generation values. As I performed this, I created three statistical measures for each climate parameter, namely *minimum*, *maximum* and *average*, which measured the monthly minimum, maximum and average values of each climate variable. The data processing pipeline is explained in the Appendix which goes into greater detail of showcasing the various steps involved. It is also worthwhile to

note that my data aggregation exercise is functional, thereby, easily able to create new statistical summary measures like *quartiles*, *Inter Quartile Range (IQR)*, *median*, *mode*, etc., which can also be useful monthly aggregate climate variables for future studies. Table 2.1 shows the relevant NARR climate variables used in my study. The final dataset, therefore, has 11 NARR variables with 3 statistical measures each, giving a total of 33 predictor climate variables.

No.	Variable Name	Units
1	convective precipitation	kg/m^2
2	air surface temperature	$kelvin$
3	albedo	%
4	total precipitation	kg/m^2
5	dew point temperature	$kelvin$
6	potential surface temperature	$kelvin$
7	relative humidity	%
8	total cloud cover	%
9	eastward windspeed	m/s
10	northward windspeed	m/s
11	total windspeed	m/s

Table 2.1. Observational climate data variables of NARR dataset.

2.2 Prediction Modeling Objectives

Using the dataset (described above) and the statistical learning techniques (discussed in the next section), generation prediction models are first trained and their performance tested. For these, a total of 34 variables comprise the set of predictor variables. 33 observational climate data variables and 1 plant capacity variable from EIA dataset are the predictors and monthly generation of corresponding power plants is the response variable. Once the best performing technique among these is discerned, a generalized prediction model is developed with the same predictor variables. However, for this latter model, *CUF* (discussed further in the Model Building Approach section below) is the response variable.

2.3 Statistical Learning Techniques

The machine learning algorithms used throughout this study can be classified under supervised learning. This is a statistical learning method that has been widely used in

the areas of risk and resilience analysis [45–47]. These algorithms predict target variables given a set of predictors. Their aim is to approximate the response variable while keeping the expected error at a minimum. In this paper, five such predictive modeling algorithms were tested and the best performing model was selected from Ordinary Least Squares Linear Regression (OLS), Bayesian Linear Regression (BLR), Classification And Regression Tree ($CART$), Gradient Boosting Machine (GBM) and Random Forest (RF).

As I have multiple explanatory variables in this study, multiple linear regression models are a good starting point to quantify the relationship between dependent and independent variables. It builds upon linear regression and assists in working with multiple independent variables. Each independent variable and its effect on the target variable can be observed [49]. Ordinary Least squares (OLS) method is generally used in defining the objective function of the linear regression technique and to estimate the β values which quantify the impact of each independent variable on the response variable. This method minimizes the difference between actual and predicted values of dependent variables [50]. Unlike OLS regression, Bayesian Linear Regression (BLR) method utilizes the prior knowledge of the variables into how β values can be estimated [48]. Therefore, the fundamental distinction here is how OLS treats model parameters as fixed but unknown quantities and finds out point estimates by minimizing the sum of squared residuals, whereas BLR considers these parameters as random distributions which are updated and accurate model estimate distributions can be obtained by using the posterior variable distributions from the train data. In this study, I have chosen linear regression models to benchmark the generation prediction models and effectively identify and quantify any linear interaction terms between the many climate variables and the generation response variable. Since OLS and BLR represent the above mentioned two distinct ways of quantifying their model parameters and achieving optimized performance, I included both these methods in the generation predictive model building process to show if the prior knowledge and posterior variable distributions carry any weight in improving/decreasing the consequent linear regression performance.

Although both OLS and BLR are frameworks of realizing a linear regression model, it is to be noted here that *"linear"* regression in itself can have polynomial or exponential/logarithmic and other functional transformation terms of the predictor variables. However, in this study, I leverage a dataset comprising the generation from a portfolio of 3 renewable sources. Even as a majority of previous and current research discussed above is focused in single energy source modeling domain, it is quite difficult in establishing

these physics-based/rule-based polynomial and non-linear effect terms of these climate variables on renewable generation based on purely linear regression techniques. Hence, only linear effects of my study dataset climate variables are used to build both OLS and BLR and then a comparitive analysis is conducted with other tree-based models (discussed below) with proven better and easier non-linear effects modeling capability.

$CART$ is a well known decision tree algorithm [43] in machine learning which can be used for both classification and regression. It works by recursively partitioning the feature space into small parts based on predictor variables values and create a relationship between them. By such a process, CART can increase the information gain while reducing the noise [51]. Such node to node connection looks like tree branches and hence the name. $CART$ offers good interpretability along with the power to handle non-linear relationships between predictor and response variables. In this study, I chose $CART$ modeling technique as it falls midway between the high interpretable but poor non-linear model performance of OLS/BLR and the lesser interpretable but high non-linear model performance of ensemble learning techniques. By employing $CART$, I can see the shifts in model performance of the generation prediction models as I move from linear to non-linear capable modeling domain. And with $CART$ model in my study, I can observe how better or worse a decision tree technique can model and quantify the various climate variables interactions with the response variable as well as perform compared to the linear regression methods.

Tree-based ensemble learning methods, such as Gradient Boosting Machines (GBM) and Random Forest (RF) solve the problem of model performance and prediction by combining multiple sub-optimal performing models, also called as weak learners, into generating a strong predictor. GBM achieves this by building a sequence of decision trees in an iterative manner [44]. In order to do so, multiple weak learners are trained and their residuals are collected which then is used to train the next set and hence at each iteration, the model can predict the errors of earlier ones and there increasing the accuracy of prediction overall [51]. Random Forest, on the other hand, uses a data resampling and simultaneous model training method that can build multiple de-correlated decision trees and combining these outputs by averaging, in case of regression study like mine, to obtain a single strong prediction value [42]. Hence, it is a method which can produce more accurate predictions. Apart from better and more accurate predictions, RF can also be robust to overfitting due to how it is constructed, and has great ability to capture non-linear data relationship as well [45, 51]. Since both GBM and RF can capture the non-linear climate-generation interaction terms and use these to give models

with better predictive power, they are used in my study to demonstrate the potential of developing generalized prediction models which can give reasonably good performance across a wide range of stratified test data all the while showing better performance than both linear regression and also conventional $CART$ decision tree.

2.4 Model Building Approach

This study uses the multidimensional dataset that is obtained by combining the EIA and NARR datasets. Each row gives a snapshot of a particular renewable plant's surrounding climate variables and its generation and operational capacity for a particular month in the study time period, with monthly generation as the response variable. Leveraging this rich and high-resolution data, I employ three types of statistical learning techniques: 1. Linear regression, 2. Decision Tree, and 3. Tree-based ensemble learning. Within linear regression category, I have used two models, Ordinary Least Squares (OLS) and Bayesian Linear Regression (BLR). For decision tree, I used the Classification and Regression Tree (CART) [43] and for the third type, I have used two models, the Gradient Boosting Machine (GBM) [44] and Random Forest (RF) [42] techniques. The modeling process is outlined in Figure 2.1.

First, the dataset was split *75%/25%*, the former for training the 5 models and the later to test the best performing model. This train-test-split was stratified across time and energy source variables to ensure that any data asymmetries and seasonal trend captures

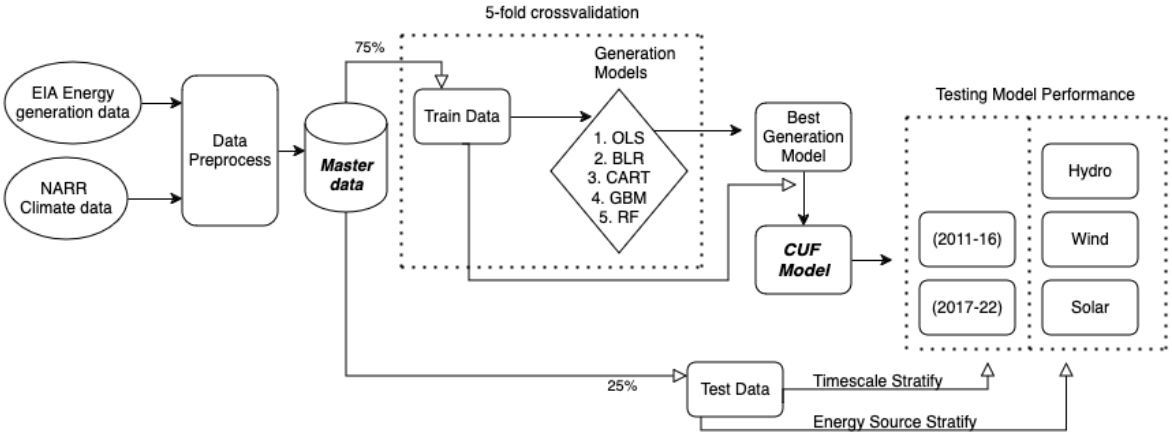


Figure 2.1. Model building flowchart used in my study. Within models, OLS and BLR are the two linear regression models. $CART$ is the decision tree. GBM and RF are the two tree-based ensemble learning methods.

are taken care of. With *generation* as the response variable and *capacity* and *NARR climate variables* as predictors, 5-fold cross validation was conducted and the models were trained. Since the response variables are quantitative and numerical, regression suitable performance metrics like Coefficient of Determination (R^2) and Normalized RMSE ($NRMSE$) were used.

Following this benchmarking of generation prediction models, I observe the best performing technique, presumably with the highest ability to capture not only linear but also non-linear effects of climate variables on generation, and it is selected as the optimum technique to build the hypothesized generalized model with CUF as the response variable. To account for the climate-sensitive portion of generation, I have introduced the Capacity Utilization Factor (CUF) as the new response variable, defined in Equation 2.1.

$$CUF = \frac{Generation\,(in\,MWh)}{[Capacity\,(in\,MW) \times \#\,hours]} \tag{2.1}$$

Using CUF as the response variable instead of generation has two objectives: (1) To eliminate the linear dependence of capacity on generation, if any, and (2) To establish that the performance of CUF prediction models is purely accounting for the non-linear climate variable impacts on CUF which serves the purpose of measuring operational *efficiency* of these renewable energy generators. I used the same train data and the best performing model to train a CUF prediction model. To demonstrate the potential of the resulting model to be termed as *generalized*, the test dataset is further stratified across timescale and energy source dimensions. For this, I performed two separate stratification actions on the test data: 1. Two subsets of test data, one having the entries from (2011-16) time period only and the other containing exclusively (2017-22) period, and 2. Three independent test subsets, one for each energy source ($Hydro$, $Wind$ and $Solar$) were created.

Chapter 3
Results

To quantify the non-linear effects of climate variables on renewable power plant generation, I first established the predictive power of linear regression models by taking climate variables and capacity as independent variables and monthly generation as the dependent variable. By looking at the improvement in performance from linear regression to decision tree to ensemble learning models, I show that this increment is the result of tree-based models being able to capture the non-linear effects of climate variables with generation. Using the best performing technique in generation prediction, I then developed the CUF prediction model and tested its performance in test data stratified across time and 3 renewable sources. This best performing CUF prediction model is then compared with source-specific model performance to show the potential of developing geo-agnostic, source-agnostic models aimed towards a multi-component integrated assessment pathway for long-term planning of climate change impacts on renewable energy systems.

3.1 Generation Prediction Models

3.1.1 Cross validation

Figure 3.1 shows the boxplots comparison of R^2 values across the 5 generation prediction models and Figure 3.2 shows the boxplots comparing $NRMSE$ values across the same models. The validation scores are practically same as the test data performance metrics presented in Figure 3.3 which shows that the models are consistent in cross validation. Here, the worst performers are the two linear models whose R^2 values hover around the same benchmark of 0.77-0.78 and the $NRMSE$ values around 0.5. This corraborates that the linear models are only effectively capturing the capacity-generation relationship and the climate variables are not adding any substantive predictive power as climate

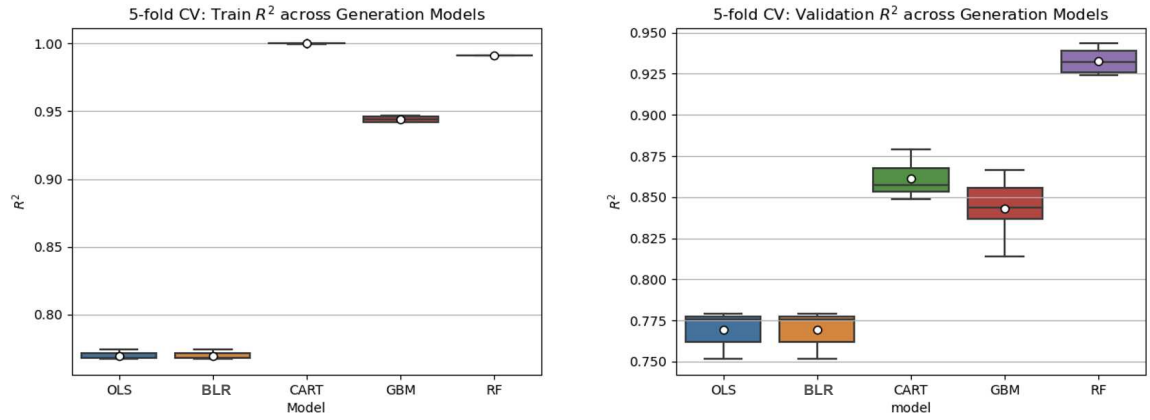


Figure 3.1. Boxplots comparing 5-fold cross validation R^2 across generation prediction models. The right plot is the boxplot of the validation hold-out subset within each fold and the left boxplot is from the R^2 values of the remaining train data in each fold.

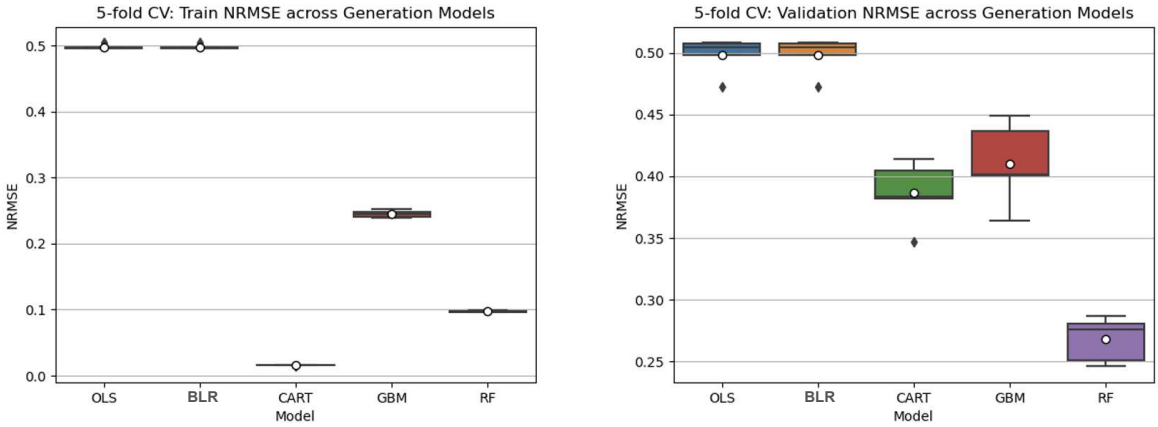


Figure 3.2. Boxplots comparing 5-fold cross validation $NRMSE$ across generation prediction models. The right plot is the boxplot of the validation hold-out subset within each fold and the left boxplot is constructed from the $NRMSE$ values of the remaining train data in each fold.

interactions with generation are non-linear. Thus, the betterment seen in the performance and predictive power is due to using tree-based ensemble learning methods which are adept in modeling those non-linear functions [36–38, 45]. The resulting broader trend of increasing R^2 and decreasing $NRMSE$ can be observed within the comparison of boxplots of the validation scores from linear to non-linear modeling techniques. The means and medians of the R^2 validation boxplots for $CART$, GBM and RF also are around their corresponding test data performance values, 0.85-0.86 for $CART$, 0.84-0.85 for GBM and 0.93-0.94 for RF. The same can also be observed to be true for validation scores of these tree-based models. As observed in these cross validation scores and

plots, RF is the best performer by a long shot with the validation R^2 values ranging between 0.925-0.945 and $NRMSE$ ranging from 0.245-0.285. This is also corroborated by observing that RF, although being marginally worse off than $CART$ in performance across the train folds of the crossvalidation, is more resilient by displaying markedly better performance in the hold-out validation fold. This is line with the potential of opting an ensemble tree approach being better than the predictive power and overfitting error that can be in the case of having a single decision tree prediction.

3.1.2 Train/Test Performance comparison

For the set of models predicting generation, monthly generation was modeled using capacity and climate variables as explanatory variables. Figure 3.3 shows the comparative analysis of the performance metrics, R^2 in the left plot and $NRMSE$ in the right plot, within each plot progressing from linear regression models to decision tree to ensemble learning models as I move from left to right. For OLS and BLR, the R^2 values for train and test data are 0.78 and 0.77 respectively, showing here again that the capacity-generation linear relationship gives us the "benchmark" modeling predictive power of explaining about 77-78% of the variance within both train and test dataset. This is also evident from looking at the right plot with the linear models having NRMSE values of 0.47 and 0.5 for train and test data respectively, showing that the linear models benchmark the performance of generation prediction models in terms of both metrics. There is also a visible trend of increment in R^2 and decrease in $NRMSE$ as I move along the x-axis of these plots from linear models to decision tree to tree-based ensemble learning

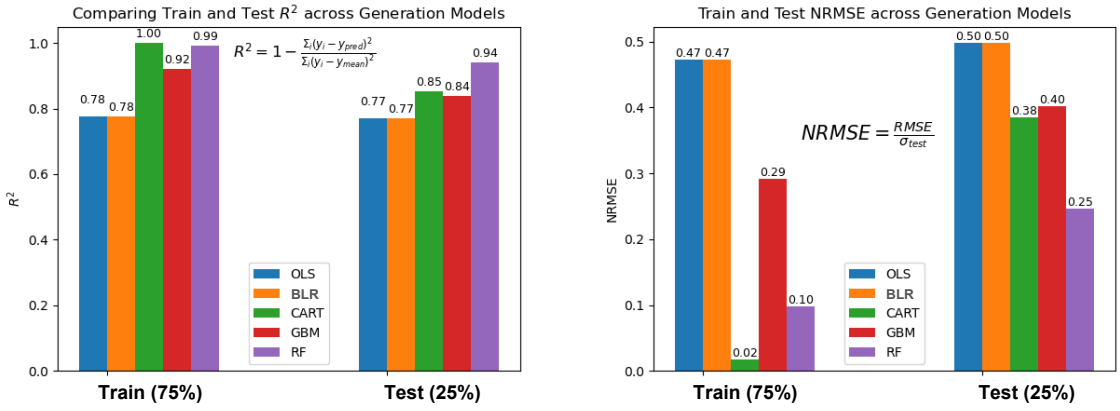


Figure 3.3. Comparing train and test performance R^2 and $NRMSE$ values across generation prediction models.

methods. Decision tree model $CART$ gives improved performance compared to the two linear regression models. This is observed in the increased R^2 to 0.85 and decreased $NRMSE$ to 0.38 for $CART$ in test data performance. GBM showed marginally worse off values than $CART$ in test data, with R^2 and $NRMSE$ of 0.84 and 0.4.

Comparing the test data performance metrics, it shows that employing modeling techniques which can capture the non-linear interactions of climate variables results in better R^2 and less erroneous models. The best performing model in terms of all three performance metrics is Random Forest, with R^2 values of 0.99 and 0.94 and $NRMSE$ values 0.1 and 0.25 for the train and test data. This performance by RF is the highest in terms of R^2 and least in terms of $NRMSE$ within the 5 generation prediction models in comparison. Therefore, an improved performance with an the increase in R^2 of 0.17 and a corresponding decrease in $NRMSE$ of 0.25 is demonstrated by using tree-based ensemble learning compared to linear regression.

3.2 CUF Prediction Models

3.2.1 Train/Test Performance comparison

As Figure 3.3 established that capacity-generation linear relationship had a benchmark performance value, CUF prediction models were developed with the objective to investigate if there is any latent linear interactions between climate variable and the capacity-normalized generation, also known as *CUF*. If all climate-CUF interactions are non-linear nature, then the predictive power of the linear models should take a severe dip and their R^2 value will go down close to zero. Similarly, the linear models RMSE and NRMSE values will still remain the highest, and considerably higher than the non-linear ensemble learning models.

By observing Figure 3.4, it is evident that all the above expected results are true. Figure 3.4 shows the comparative bar graphs of the two performance metrics, R^2 in the left plot and NRMSE in the right, with each plot axis moving from linear models to decision tree to ensemble learning models. For the linear CUF prediction models, the R^2 values for both train and test data are pretty much close to 0, at a meagre 0.05, showing that there is little to none linear effects in the climate-CUF relationship. Consequently, the linear models are the ones with the highest $NRMSE$ values (0.98 for both train and test data). The previously observed trend of increasing R^2 and reducing $NRMSE$ as I move from linear predictive models to non-linear ensemble learning models can

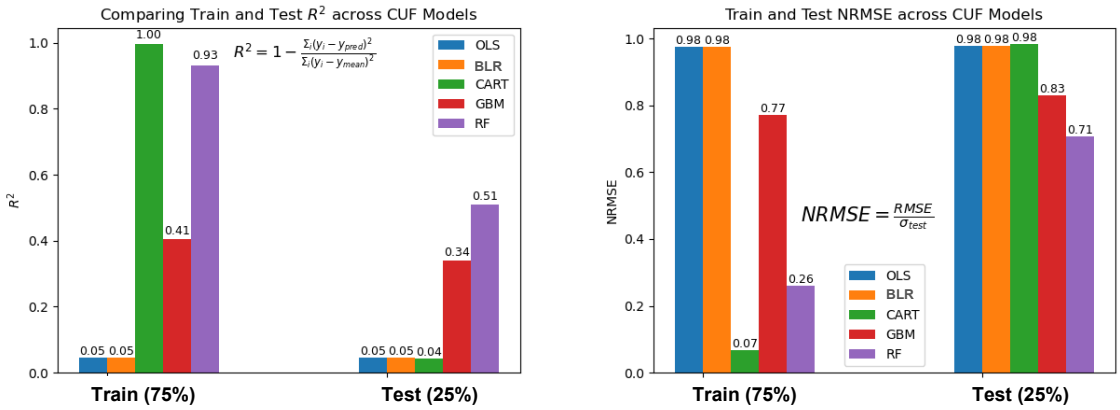


Figure 3.4. Comparing train and test performance R^2 and $NRMSE$ across CUF prediction models.

also be seen in CUF prediction models. This is a confirmation that taking CUF as the response variable has only eliminated any linear terms on the explanatory variables side of the functional relationship and this change in the response variable has enabled the identification of predictive models which can capture the non-linear effects between the explanatory and response variables. The $CART$ decision tree model gives the same R^2 value as linear models (close to 0) for test data and even the test NRMSE values are in the same range as the values of linear models. This shows that using a single decision tree may not produce the expected predictive power in my study, even though decision trees can have the ability to model non-linear functions to a good extent [45]. The results of CUF prediction models demonstrates that RF can leverage the principle of ensemble learning and produce substantially good predictions. This is evident in the plots that RF is the best performer even when CUF is the response variable, with test data performance values of 0.51 and 0.71 for R^2 and NRMSE respectively.

3.3 Best performing CUF Prediction: A Generalized model

Once the best performing model, RF here, is observed, this CUF prediction model has been tested across multiple independently stratified test datasets. This is done to demonstrate the ability of RF to be a *generalized* model and have a good model performance across a wide range of possible testing datasets.

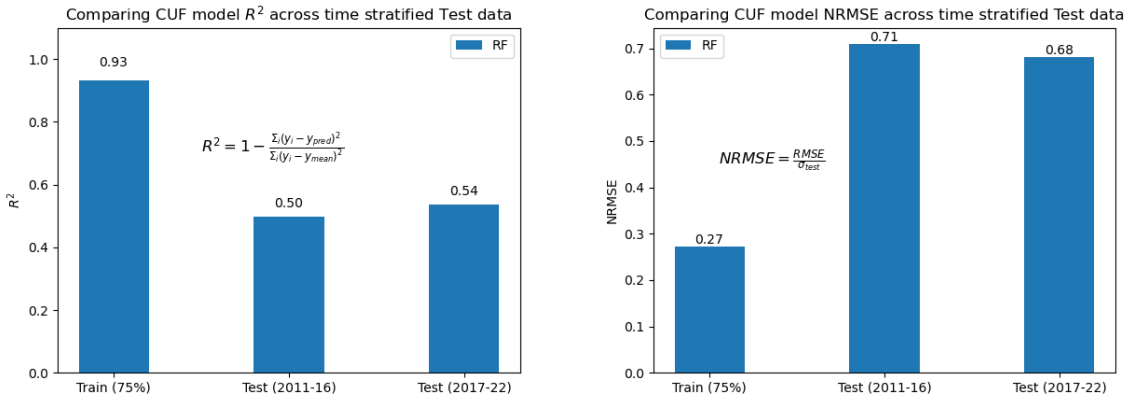


Figure 3.5. Comparing test performance R^2 and NRMSE of best performing CUF prediction model (RF) across temporal stratified test data. First test data subset is the period (2011-16) and the Second subset is (2017-22) time period.

3.3.1 Testing across time-stratified data

Figure 3.5 illustrate the performance of RF CUF prediction models using test data that is stratified across timescales. The left plot in this figure is for R^2 performance metric and the right plot in the figure is for $NRMSE$ respectively. The middle bar in these figures shows the performance on using the test data subset of 2011-16 time period whereas the right most bars are for the 2017-22 test data period. Comparing the middle and right bars of the left plot in Figure 3.5 with RF performance values in the left plot of Figure 3.4 validates that the R^2 values of the CUF prediction models are practically the same for the total test data (0.51) as well as the 2011-16 test data subset (0.5) and the 2017-22 stratified test data (0.54). The same observation can be made for $NRMSE$ metrics too by comparing the right plot in Figure 3.5 and the right plot in Figure 3.4 with 0.71 for total test data, 0.71 for 2011-16 test subset and 0.68 for 2017-22 time period test data. These figures illustrate that the CUF prediction model developed using the best performing RF technique has generalized and consistent performance across different and distinct test data timescales.

3.3.2 Testing across source-stratified data

For the hypothesized generalized model to have consistent predictive power across the threee renewable energy sources, Hydro, Wind and Solar, the originally 25% hold-out data put away from model training phase was further divided into 3 distinct subsets, one for each energy source. Using these energy source stratified subsets, the potential of

the generalized RF CUF prediction model is tested by comparing with the respective performance of 3 source-specific models trained on the train data subset of corresponding energy source.

3.3.2.1 Source-specific CUF Prediction Models

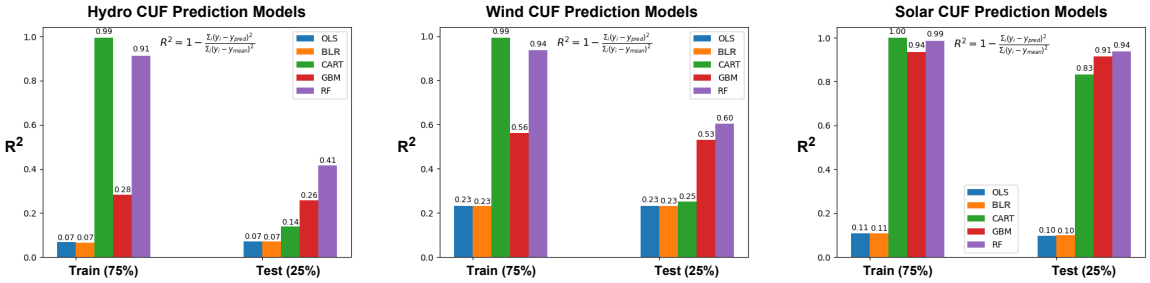


Figure 3.6. Comparing R^2 performance of source-specific CUF prediction models. *RF* is the best performing modeling technique for all three energy sources.

Figure 3.6 shows the development of source-specific CUF prediction models for the 3 energy sources, Hydro, Wind and Solar, using the same linear vs non-linear statistical learning techniques approach. All the models whose performances are illustrated in this figure have been trained using the same 5-fold crossvalidation model building framework as previously discussed in Chapter 2, Figure 2.1. The left-most plot is the comparison of CUF prediction models trained on the Hydro subset data of the 75% train data and tested on the Hydro stratified subset of the 25% hold-out data. Similarly, the center plot is comparing Wind CUF prediction models trained on the Wind subset of the 75% train data and tested on the Wind energy source stratified subset of the 25% hold-out test data. And the right plot is for the Solar CUF prediction models trained and tested on the Solar subsets of the same datasets respectively.

Figure 3.6 shows that *RF* is the best performing modeling technique for all three energy sources, with R^2 of 0.91 and 0.41 for Hydro CUF model performance on train and test data, for Wind CUF prediction model with R^2 of 0.94 and 0.6 on wind subset train and test data, and Solar CUF prediction model having R^2 of 0.99 and 0.94 on solar subset train and test data respectively. For the next section where I investigate how well the generalized model performs in energy source stratified test data, these RF source-specific models are used for comparison.

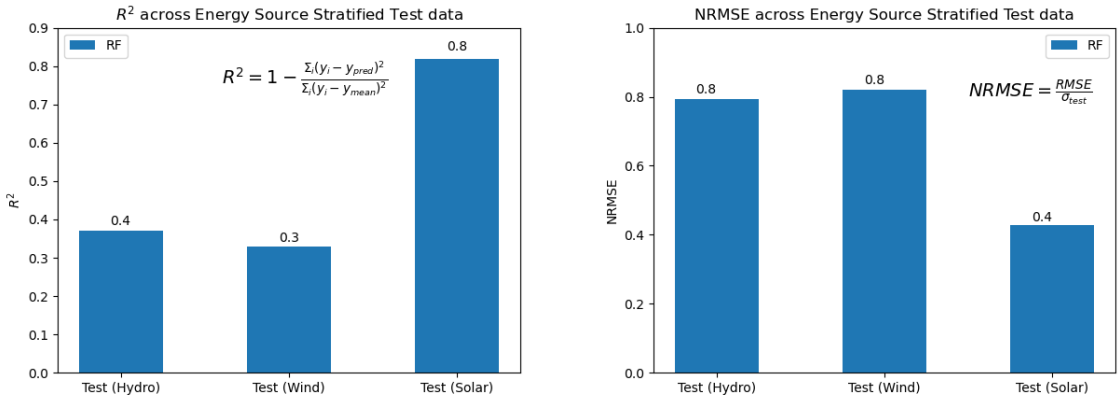


Figure 3.7. Test performance R^2 and $NRMSE$ of best performing CUF prediction model (RF) across energy source stratified test data

3.3.2.2 Comparing Generalized model with source-specific models

For Figure 3.7, I performed test data stratification using energy source variable. There were three subsets of the test data, one for each renewable source of energy in the study dataset ($Hydro$, $Wind$ and $Solar$), were obtained and the performance of CUF prediction models is shown in Figure 3.7. The left plot shows the comparitive analysis of each energy source stratified test data, with energy sources on x-axis and R^2 values on the y-axis. The right plot shows the same but with $NRMSE$ on the y-axis. One observation is that the model performance is very much dependent on the type of energy source. For Hydro test data, the R^2 and $NRMSE$ values are 0.4 and 0.8 respectively. These values are very close to the model performance on Wind test data, with the latter having R^2 of 0.3 and $NRMSE$ of 0.8. Solar test data is where the model gives the best performance in terms of both metrics, with R^2 value of 0.8 and $NRMSE$ value 0.4.

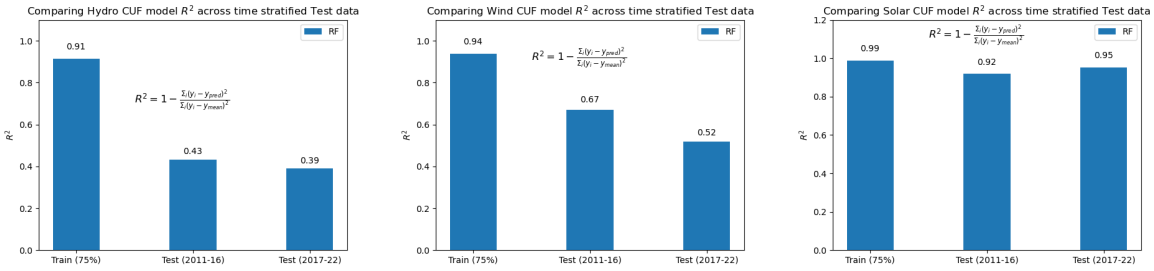


Figure 3.8. Test performance R^2 and $NRMSE$ of source-specific best performing CUF prediction model.

Figure 3.8 illustrates the performance of the best source-specific model, RF in all

three cases, on time-stratified test data. Observing the Figure 3.7 and Figure 3.8, it is evident that the *generalized* CUF model has performance that is consistent and close to the Hydro and Solar CUF prediction models. With R^2 of 0.4 and 0.8 for Hydro and Solar stratified test subsets respectively, the generalized model is able to perform on par with both Hydro (test R^2 of 0.43 and 0.39) as well as Solar (test R^2 of 0.92 and 0.95) source-specific RF models. Wind CUF prediction model outperforms the generalized model by some distance with R^2 of 0.67 and 0.52 for the former and R^2 value of only 0.3 for the latter.

3.4 Discussion

Looking at the overall picture, it can be observed that RF has the best model performance across both time as well as energy source stratified test data within the EIA-NARR custom dataset I developed. Although the performance values by even this best performing RF are not great in absolute terms, this study is aimed at contexualizing and understanding the performance of this *generalized* model with respect to source-specific models to have a potential of being a geo-agnostic, source-agnostic model. With R^2 and $NRMSE$ values 0.51 and 0.71 of the hypothesized generalized model, it is comparable with hydro and wind CUF prediction models, although still much lesser than the solar CUF model performance of R^2 0.94. However, observing the similar trend within each energy source stratified test data shows that the generalized model performance is close to how hydro and solar source-specific models predict.

This demonstrates the ability of non-linear modeling techniques, specifically tree-based ensemble learning methods, to realize generalized prediction models which can add to our existing understanding of renewables-climate-risk research. This is, however, to be noted that such generalized modeling framework is in the need of having efforts towards medium-term and long-term understanding of how climate change impacts can affect our increasingly diversifying energy systems portfolio. Existing research and literature in conducting source-specific modeling and analysis of climate impacting our current renewables stock is very important in driving our current-term and short-term optimization of energy system operations.

This study and research work is aimed at envisioning a future, in line with the trends of rising renewables in the mix, with the explicit research need to look at common data pipeling and modeling frameworks in developing data-driven models which can be geo-agnostic as well as source-agnostic.

3.5 Study Limitations and Future Work

This study currently only uses publicly available USA datasets, which limits the geographical extent of the renewable power plant data. But the open-source nature of this research and data pipelines is explicitly to enable widespread collaboration on this research and enable scaling up of data and modeling framworks. By this research, my objective of showing the potential of a generalized CUF prediction model has been demonstrated, albeit within the context of USA renewables. This is why my custom dataset has been open-sourced along with the data-pipeling schema, to enable multidisciplinary and cross-country collaboration and append the current data with other countries' data.

Additionally, my study has focused on the three major renewable energy sources which I have to found to be most contingent on ambient climatic conditions i.e., hydro, solar and wind. To fully realize geo-agnostic and energy source-agnostic prediction models, we need to include other energy systems within my analyses to better understand the potential strengths as well as limitations of the RF-based generalized prediction model. As we have already observed, model performance is linked to the type of renewable source involved in the operations of particular stakeholders.

The current choice of using a monthly time series data suffers with two limitations. A month-level granularity of data and modeling framework is not of effective use when it comes to now-term and short-term planning and operations of energy systems. So, this study approach is really intended to augment our understanding and planning efforts of climate change impacts on energy systems in the long-term domain. This is majorly in line with our growing need to develop models in place which can be used to inform long-term impact assessment and resilience. Second, a monthly time series data modeling framework is not great for especially understanding hydrological systems [63]. This may also be evident in my results that hydro CUF as well as generalized model performance on hydro stratified test data is so low, with around 0.4 value of R^2 for both.

Another limitation mainly arises from the need to adequately account for the spatial and temporal covariances and correlations here. For a long-term holistic projection of a portfolio-level climate change impacts, it is critical to look at geographical variations within the dataset and decode any correlations of climate and generation variables across different locations in a snapshot of time. Similarly, the auto-correlation aspect of the portfolio's generation and climate variables needs to be explored. By opting for crossvalidation to build train models, this temporal aspect of dependency trends is currently missing.

Chapter 4
Summary and Conclusions

This study aimed to understand how large, multidimensional and robust datasets can be leveraged in producing generalized predictive models which can give good predictions in a variety of stratified test datasets. In this study, publicly available data on USA energy systems portfolio and latitude-longitude grid climate variables data were used to create a novel and rich dataset. Using this custom dataset, the performance of linear regression models was shown to be the benchmark of capturing the renewable plant capacity-generation linear relationship. Tree-based ensemble learning methods which have a good potential to capture the non-linear climate-generation interactions showed an improvement in performance in 2 metrics, R^2 and NRMSE. The least performing among the 5 models were the linear regression models and the best performer is the Random Forest. Using RF modeling, a CUF prediction model was also developed and train/test performance metrics were obtained to show that this captured the non-linear ways of hyper-local climate variables interacting with renewable plants' operational efficiency. To demonstrate the potential of developing generalized models, the performance of the tree-based ensemble learning model in stratified test data was studied. Test data was divided into 2 subsets in time dimension (2011-16 and 2017-22 periods) and 3 subsets, one for each renewable energy source of hydro, wind and solar. This study demonstrates that the hypothesized generalized model can have consistent and comparable model performance with respect to source-specific models, at least for Hydro and Solar in the current shape. This breakthrough has huge potential to address global energy systems trade-offs between geo-specific and source-specific models and enable fast and scalable renewables planning as well as operational expansions. This research is intended to address some of these critical research gaps in our existing literature on renewables-climate-risk.

Bibliography

[1] Craig, M. T., Cohen, S., Macknick, J., Draxl, C., Guerra, O. J., Sengupta, M., Brancucci, C. (2018). A review of the potential impacts of climate change on bulk power system planning and operations in the United States. Renewable and Sustainable Energy Reviews, 98, 255–267.

[2] Dosio, A., Mentaschi, L., Fischer, E. M., & Wyser, K. (2018). Extreme heat waves under 1.5 C and 2 C global warming. Environmental Research Letters, 13(5), 054006.

[3] Ridder, N. N., Ukkola, A. M., Pitman, A. J., & Perkins-Kirkpatrick, S. E. (2022). Increased occurrence of high impact compound events under climate change. Npj Climate and Atmospheric Science, 5(1), 3.

[4] National Academies of Sciences, Division on Earth, Life Studies, Board on Atmospheric Sciences, Committee on Extreme Weather Events, & Climate Change Attribution. (2016). Attribution of extreme weather events in the context of climate change. National Academies Press.

[5] Yalew, S. G., van Vliet, M. T. H., Gernaat, D. E., Ludwig, F., Miara, A., Park, C., Others. (2020). Impacts of climate change on energy systems in global and regional scenarios. Nature Energy, 5(10), 794–802.

[6] Van Vliet, M. T. H., Yearsley, J. R., Ludwig, F., Vögele, S., Lettenmaier, D. P., & Kabat, P. (2012). Vulnerability of US and European electricity supply to climate change. Nature Climate Change, 2(9), 676–681.

[7] Global Energy & CO2 Status Report 2019 – Analysis - IEA — iea.org. (2019). Retrieved from https://www.iea.org/reports/global-energy-co2-status-report-2019

[8] Allen, M., Dube, O. P., Solecki, W., Aragón-Durand, F., Cramer, W., Humphreys, S., Others. (2018). Global warming of 1.5 C. An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.

[9] Cronin, J., Anandarajah, G., & Dessens, O. (2018). Climate change impacts on the energy system: a review of trends and gaps. Climatic Change, 151(2), 79–93.

[10] Craig, M. T., Wohland, J., Stoop, L. P., Kies, A., Pickering, B., Bloomfield, H. C., Others. (2022). Overcoming the disconnect between energy system and climate modeling. Joule, 6(7), 1405–1417.

[11] Gernaat, D. E., de Boer, H. S., Daioglou, V., Yalew, S. G., Müller, C., & van Vuuren, D. P. (2021). Climate change impacts on renewable energy supply. Nature Climate Change, 11(2), 119–125.

[12] Schaeffer, R., Szklo, A. S., de Lucena, A. F. P., Borba, B. S. M. C., Nogueira, L. P. P., Fleming, F. P., Boulahya, M. S. (2012). Energy sector vulnerability to climate change: A review. Energy, 38(1), 1–12.

[13] Pfenninger, S., Hawkes, A., & Keirstead, J. (2014). Energy systems modeling for twenty-first century energy challenges. Renewable and Sustainable Energy Reviews, 33, 74–86.

[14] Auffhammer, M., Baylis, P., & Hausman, C. H. (2017). Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States. Proceedings of the National Academy of Sciences, 114(8), 1886–1891.

[15] Sathaye, J. A., Dale, L. L., Larsen, P. H., Fitts, G. A., Koy, K., Lewis, S. M., & de Lucena, A. F. P. (2013). Estimating impacts of warming temperatures on California's electricity system. Global Environmental Change, 23(2), 499–511.

[16] Van Vliet, M. T. H., Yearsley, J. R., Ludwig, F., Vögele, S., Lettenmaier, D. P., & Kabat, P. (2012). Vulnerability of US and European electricity supply to climate change. Nature Climate Change, 2(9), 676–681.

[17] Obringer, R., Nateghi, R., Maia-Silva, D., Mukherjee, S., Cr, V., McRoberts, D. B., & Kumar, R. (2022). Implications of increasing household air conditioning use across the United States under a warming climate. Earth's Future, 10(1), e2021EF002434.

[18] McFarland, J., Zhou, Y., Clarke, L., Sullivan, P., Colman, J., Jaglom, W. S., Others. (2015). Impacts of rising air temperatures and emissions mitigation on electricity demand and supply in the United States: a multi-model comparison. Climatic Change, 131, 111–125.

[19] Li, D. H. W., Yang, L., & Lam, J. C. (2012). Impact of climate change on energy use in the built environment in different climate zones–a review. Energy, 42(1), 103–112.

[20] Isaac, M., & Van Vuuren, D. P. (2009). Modeling global residential sector energy demand for heating and air conditioning in the context of climate change. Energy Policy, 37(2), 507–521.

[21] Chandramowli, S. N., & Felder, F. A. (2014). Impact of climate change on electricity systems and markets–A review of models and forecasts. Sustainable Energy Technologies and Assessments, 5, 62–74.

[22] on Climate Change (IPCC), I. P. (2014). Climate Change 2014 – Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects: Working Group II Contribution to the IPCC Fifth Assessment Report. Cambridge University Press.

[23] Van Vliet, Michelle T. H., Yearsley, J. R., Ludwig, F., Vögele, S., Lettenmaier, D. P., & Kabat, P. (2012). Vulnerability of US and European electricity supply to climate change. Nature Climate Change, 2(9), 676–681.

[24] Karnauskas, K. B., Lundquist, J. K., & Zhang, L. (2018). Southward shift of the global wind energy resource under high carbon dioxide emissions. Nature Geoscience, 11(1), 38–43.

[25] Wild, M., Folini, D., Henschel, F., Fischer, N., & Müller, B. (2015). Projections of long-term changes in solar radiation based on CMIP5 climate models and their influence on energy yields of photovoltaic systems. Solar Energy, 116, 12–24.

[26] Stanton, M. C. B., Dessai, S., & Paavola, J. (2016). A systematic review of the impacts of climate variability and change on electricity systems in Europe. Energy, 109, 1148–1159.

[27] Bartos, M. D., & Chester, M. V. (2015). Impacts of climate change on electric power supply in the Western United States. Nature Climate Change, 5(8), 748–752.

[28] Van Vliet, M. T. H., Van Beek, L. P. H., Eisner, S., Flörke, M., Wada, Y., & Bierkens, M. F. P. (2016). Multi-model assessment of global hydropower and cooling water discharge potential under climate change. Global Environmental Change, 40, 156–170.

[29] Hamududu, B., & Killingtveit, A. (2017). Assessing climate change impacts on global hydropower. In Climate change and the future of sustainability (pp. 109–132). Apple Academic Press.

[30] Fant, C., Schlosser, C. A., & Strzepek, K. (2016). The impact of climate change on wind and solar resources in southern Africa. Applied Energy, 161, 556–564.

[31] Jerez, S., Tobin, I., Vautard, R., Montávez, J. P., López-Romero, J. M., Thais, F., Others. (2015). The impact of climate change on photovoltaic power generation in Europe. Nature Communications, 6(1), 10014.

[32] Folini, D., Dallafior, T. N., Hakuba, M. Z., & Wild, M. (2017). Trends of surface solar radiation in unforced CMIP5 simulations. Journal of Geophysical Research: Atmospheres, 122(1), 469–484.

[33] Johnson, D. L., & Erhardt, R. J. (2016). Projected impacts of climate change on wind energy density in the United States. Renewable Energy, 85, 66–73.

[34] Haupt, S. E., Copeland, J., Cheng, W. Y. Y., Zhang, Y., Ammann, C., & Sullivan, P. (2016). A method to assess the wind and solar resource and to quantify interannual variability over the United States under current and projected future climate. Journal of Applied Meteorology and Climatology, 55(2), 345–363.

[35] Kostevica, V., & Dzikevics, M. (2023). Bibliometric Analysis of the Climate Change Impact on Energy Systems. Environmental and Climate Technologies, 27(1), 950–963.

[36] Galelli, S., & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. Water Resources Research, 49(7), 4295–4310.

[37] Obringer, R., Mukherjee, S., & Nateghi, R. (2020). Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework. Applied Energy, 262, 114419.

[38] Mukherjee, S., & Nateghi, R. (2017). Climate sensitivity of end-use electricity consumption in the built environment: an application to the state of Florida, United States. Energy, 128, 688–700.

[39] Form EIA-860 detailed data with previous form data (EIA-860A/860B) — eia.gov. (n.d.). Retrieved from https://www.eia.gov/electricity/data/eia860/

[40] Form EIA-923 detailed data with previous form data (EIA-906/920) - U.S. Energy Information Administration (EIA) — eia.gov. (n.d.). Retrieved from https://www.eia.gov/electricity/data/eia923/

[41] Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Others. (2006). North American regional reanalysis. Bulletin of the American Meteorological Society, 87(3), 343–360.

[42] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

[43] Breiman, L. (2017). Classification and regression trees. Routledge.

[44] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

[45] Obringer, Renee, & Nateghi, R. (2018). Predicting urban reservoir levels using statistical learning techniques. Scientific Reports, 8(1), 5164.

[46] Pezalla, S., & Obringer, R. (2023). Evaluating the household-level climate-electricity nexus across three cities through statistical learning techniques. Socio-Economic Planning Sciences, 89, 101664.

[47] Obringer, R., Kumar, R., & Nateghi, R. (2019). Analyzing the climate sensitivity of the coupled water-electricity demand nexus in the Midwestern United States. Applied Energy, 252, 113466.

[48] Barry, D. (1986). Nonparametric Bayesian Regression. The Annals of Statistics, 934–953.

[49] Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). Multivariate data analysis: A global perspective (Vol. 7). Upper Saddle River, NJ: Pearson.

[50] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. McGraw-Hill.

[51] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2). Springer.

[52] Bartos, M. D., & Chester, M. V. (2015). Impacts of climate change on electric power supply in the Western United States. Nature Climate Change, 5(8), 748–752.

[53] Craig, M. T., Losada Carreño, I., Rossol, M., Hodge, B.-M., & Brancucci, C. (2019). Effects on power system operations of potential changes in wind and solar generation potential under climate change. Environmental Research Letters, 14(3), 034014.

[54] Cronin, J., Anandarajah, G., & Dessens, O. (2018). Climate change impacts on the energy system: a review of trends and gaps. Climatic Change, 151(2), 79–93.

[55] Doss-Gollin, J., Amonkar, Y., Schmeltzer, K., & Cohan, D. (2023). Improving the Representation of Climate Risks in Long-Term Electricity Systems Planning: a Critical Review. Current Sustainable/Renewable Energy Reports, 10(4), 206–217.

[56] Khuntia, S. R., Rueda, J. L., & van Der Meijden, M. A. (2016). Forecasting the load of electrical power systems in mid-and long-term horizons: a review. IET Generation, Transmission & Distribution, 10(16), 3971–3977.

[57] Mideksa, T. K., & Kallbekken, S. (2010). The impact of climate change on the electricity market: A review. Energy Policy, 38(7), 3579–3585.

[58] Moftakhari, H., & AghaKouchak, A. (2019). Increasing exposure of energy infrastructure to compound hazards: cascading wildfires and extreme rainfall. Environmental Research Letters, 14(10), 104018.

[59] Panteli, M., & Mancarella, P. (2015). Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies. Electric Power Systems Research, 127, 259–270.

[60] Russo, M. A., Carvalho, D., Martins, N., & Monteiro, A. (2023). Future perspectives for wind and solar electricity production under high-resolution climate change scenarios. Journal of Cleaner Production, 404, 136997.

[61] Singh, D., Bekris, Y., Rogers, C. D. W., Doss-Gollin, J., Coffel, E. D., & Kalashnikov, D. A. (2024). Enhanced solar and wind potential during widespread temperature extremes across the US interconnected energy grids. Environmental Research Letters.

[62] Wang, J., Pinson, P., Chatzivasileiadis, S., Panteli, M., Strbac, G., & Terzija, V. (2022). On machine learning-based techniques for future sustainable and resilient energy systems. IEEE Transactions on Sustainable Energy, 14(2), 1230–1243.

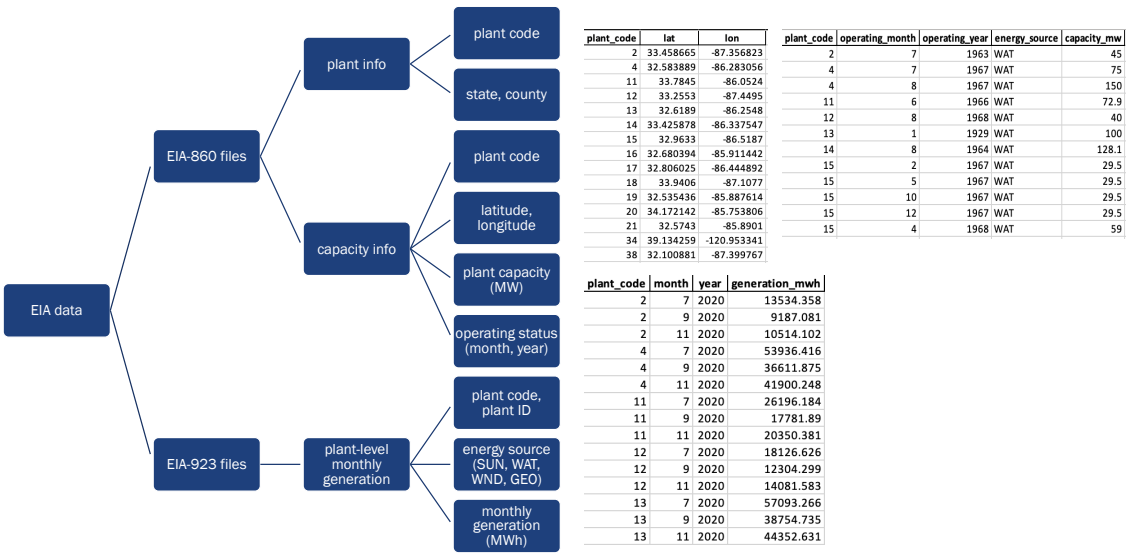
[63] Webster, M., Fisher-Vanden, K., Kumar, V., Lammers, R. B., & Perla, J. (2022). Integrated hydrological, power system and economic modelling of climate impacts on electricity demand and cost. Nature Energy, 7(2), 163–169.

[64] Perera, A. T. D., Nik, V. M., Chen, D., Scartezzini, J.-L., & Hong, T. (2020). Quantifying the impacts of climate change and extreme climate events on energy systems. Nature Energy, 5(2), 150–159.

Appendix
Data Processing Pipeline

1 EIA Renewables Portfolio Dataset

The electricity sector data across the USA was obtained from the U.S Energy Information Administration (EIA). Two distinct subsets that were used for this are the EIA-860 and EIA-923 datasets. The former contains variables which define a particular power plant within the overall portfolio across the country. These include a uniquely assigned plant code, geographical markers of the plant location such as the state, county and the latitude/longitude of the plant's location, operational plant capacity (in MW) and



plant_code	lat	lon
2	33.458665	-87.356823
4	32.583889	-86.283056
11	33.7845	-86.0524
12	33.2553	-87.4495
13	32.6189	-86.2548
14	33.425878	-86.337547
15	32.9633	-86.5187
16	32.680394	-85.911442
17	32.806025	-86.444892
18	33.9406	-87.1077
19	32.535436	-85.887614
20	34.172142	-85.753806
21	32.5743	-85.8901
34	39.134259	-120.953341
38	32.100881	-87.399767

plant_code	operating_month	operating_year	energy_source	capacity_mw
2	7	1963	WAT	45
4	7	1967	WAT	75
4	8	1967	WAT	150
11	6	1966	WAT	72.9
12	8	1968	WAT	40
13	1	1929	WAT	100
14	8	1964	WAT	128.1
15	2	1967	WAT	29.5
15	5	1967	WAT	29.5
15	10	1967	WAT	29.5
15	12	1967	WAT	29.5
15	4	1968	WAT	59

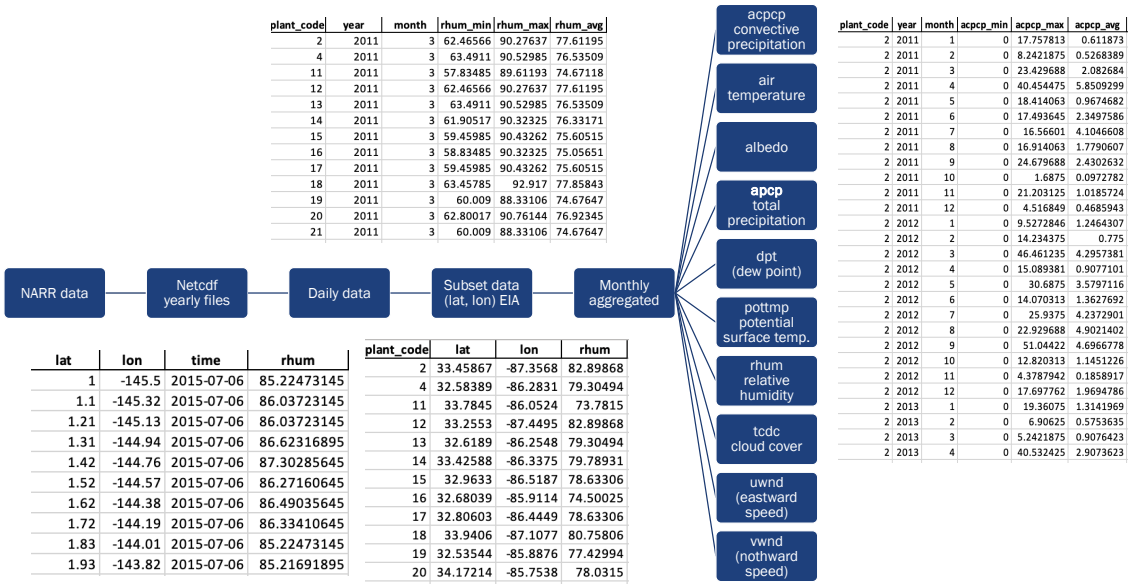
plant_code	month	year	generation_mwh
2	7	2020	13534.358
2	9	2020	9187.081
2	11	2020	10514.102
4	7	2020	53936.416
4	9	2020	36611.875
4	11	2020	41900.248
11	7	2020	26196.184
11	9	2020	17781.89
11	11	2020	20350.381
12	7	2020	18126.626
12	9	2020	12304.299
12	11	2020	14081.583
13	7	2020	57093.266
13	9	2020	38754.735
13	11	2020	44352.631

Figure A.1. EIA Datapipeline

the time since each power plant began its operations. Whereas, EIA-923 data contains the variables relevant to power plant generation like the source of energy and monthly generation (in MWh) along with the already mentioned plant code. Both these data sources are combined using the plant code variable as the primary key. Together, this data gives us the information to identify and locate any operational power plant across the United States, and analyze its monthly generation over the years.

2 NARR Historical Climate Data

The recorded historical climate data was collected from the North American Regional Reanalysis (NARR). The daily values of the NARR data were aggregated to get monthly measures. This was done to maintain parity and temporal consistency with the energy sector data which has monthly generation values. As I performed this, I created three statistical measures for each climate parameter, namely *minimum*, *maximum* and *average*, which measured the monthly minimum, maximum and average values of each climate variable.



plant_code	year	month	rhum_min	rhum_max	rhum_avg
2	2011	3	62.46566	90.27637	77.61195
4	2011	3	63.4911	90.52985	76.53509
11	2011	3	57.83485	89.61193	74.67118
12	2011	3	62.46566	90.27637	77.61195
13	2011	3	63.4911	90.52985	76.53509
14	2011	3	61.90517	90.32325	76.33171
15	2011	3	59.45985	90.43262	75.60515
16	2011	3	58.83485	90.32325	75.05651
17	2011	3	59.45985	90.43262	75.60515
18	2011	3	63.45785	92.917	77.85843
19	2011	3	60.009	88.33106	74.67647
20	2011	3	62.80017	90.76144	76.92345
21	2011	3	60.009	88.33106	74.67647

lat	lon	time	rhum
1	-145.5	2015-07-06	85.22473145
1.1	-145.32	2015-07-06	86.03723145
1.21	-145.13	2015-07-06	86.03723145
1.31	-144.94	2015-07-06	86.62316895
1.42	-144.76	2015-07-06	87.30285645
1.52	-144.57	2015-07-06	86.27160645
1.62	-144.38	2015-07-06	86.49035645
1.72	-144.19	2015-07-06	86.33410645
1.83	-144.01	2015-07-06	85.22473145
1.93	-143.82	2015-07-06	85.21691895

plant_code	lat	lon	rhum
2	33.45867	-87.3568	82.89868
4	32.58389	-86.2831	79.30494
11	33.7845	-86.0524	73.7815
12	33.2553	-87.4495	82.89868
13	32.6189	-86.2548	79.30494
14	33.42588	-86.3375	79.78931
15	32.9633	-86.5187	78.63306
16	32.68039	-85.9114	74.50025
17	32.80603	-86.4449	78.63306
18	33.9406	-87.1077	80.75806
19	32.53544	-85.8876	77.42994
20	34.17214	-85.7538	78.0315

plant_code	year	month	acpcp_min	acpcp_max	acpcp_avg
2	2011	1	0	17.757813	0.611873
2	2011	2	0	8.2421875	0.5268389
2	2011	3	0	23.429688	2.082684
2	2011	4	0	40.454475	5.8509299
2	2011	5	0	18.414063	0.9674682
2	2011	6	0	17.493645	2.3497586
2	2011	7	0	16.56601	4.1046608
2	2011	8	0	16.914063	1.7790607
2	2011	9	0	24.679688	2.4302632
2	2011	10	0	1.6875	0.0972782
2	2011	11	0	21.203125	1.0185724
2	2011	12	0	4.516849	0.4685943
2	2012	1	0	9.5272846	1.2464307
2	2012	2	0	14.234375	0.775
2	2012	3	0	46.461235	4.2957381
2	2012	4	0	15.089381	0.9077101
2	2012	5	0	30.6875	3.5797116
2	2012	6	0	14.070313	1.3627692
2	2012	7	0	25.9375	4.2372901
2	2012	8	0	22.929688	4.9021402
2	2012	9	0	51.04422	4.6966778
2	2012	10	0	12.820313	1.1451226
2	2012	11	0	4.3787942	0.1858917
2	2012	12	0	17.697762	1.9694786
2	2013	1	0	19.36075	1.3141969
2	2013	2	0	6.90625	0.5753635
2	2013	3	0	5.2421875	0.9076423
2	2013	4	0	40.532425	2.9073623

Figure A.2. NARR Datapipeline