# Predicting Virginia County Suitability for Carbon Dioxide Removal Techniques

Grace Davenport[1], Rachel Holman[1], Lillian Jarrett[1], and Hana Nur[1]

[1] University of Virginia, Charlottesville, VA 22904 USA ( e-mail: mha2rh, dnw9qk, lej4us, hrn4ch@virginia.edu).

**Carbon dioxide is the most prominent greenhouse gas in our atmosphere (Lashof & Ahuja 1990). To lessen the impact of CO2 levels in the atmosphere, various carbon dioxide removal (CDR) technologies can be deployed, such as biochar and reforestation. Where should these technologies be deployed? To answer this question, we implemented unsupervised and supervised ML models to predict suitability levels of Virginia counties across three CDR technologies. First, we implemented unsupervised ML models to cluster and label counties into four suitability levels: highly suitable, suitable, possible, and unsuitable for each CDR technology. Second, eight supervised ML models were deployed to predict each county suitability level. The model performance accuracies were very high, ranging from 82.4 - 100% with the exception of the underperforming Naive Bayes model. Temperature had a high feature importance score in all three CDR technologies. Income and agricultural land use were important features for enhanced weathering (EW) and biochar, but not reforestation. To complement these results, we performed a case study on six holdout counties. Our supervised ML models predicted that (1) Accomack, (2) Fauquier, and (3) Hanover counties were highly suitable for EW, due to their high carbon emission energy available. In the future, the scope of this project could be increased to state or even country level.**

## I. INTRODUCTION

Carbon dioxide ($CO_2$) is the most prominent greenhouse gas on Earth at 79.4%, followed by methane at 11.5% (Lashof & Ahuja 1990). When hit by infrared rays from the sun, carbon dioxide begins to vibrate, releasing heat into the atmosphere. This is called the greenhouse effect. Over time, this vibration can increase the Earth's atmospheric temperature. According to NOAA, since 1850, the Earth's temperature has increased by 0.11 degrees Fahrenheit per decade and about 2 degrees Fahrenheit total (Sweeney et al. 2015).

To mitigate this greenhouse effect and achieve a more sustainable future, various carbon removal technologies can be deployed. There are three categories of CDR technologies: (1) nature based, (2) hybrid, and (3) engineered (ClimateSeed 2024). In this study we focus on one nature based solution—reforestation—and two hybrid solutions—enhanced weathering (EW) and biochar. Reforestation, planting additional trees in forested areas, removes carbon from the atmosphere by photosynthesis. Enhanced weathering binds dissolved CO2 with ions released from rock particles, forming stable minerals (Schuiling & Krijgsman 2006). Biochar is produced when organic waste is burned in an oxygen deficient environment, stabilizing carbon dioxide from the atmosphere (Wang & Wang 2019).

Few locations are suitable for all CDR technologies, as each are highly dependent on complex interactions between bio-geophysical, socio-political, and techno-economic factors. Therefore, locations should be thoroughly investigated before implementing CDR technologies.

Previous literature has documented varying methods to ascertain location suitability for different CDR technologies. Donnison et al. determined the suitability of six UK locations for BECCS (bioenergy with carbon capture and storage) implementation with a spatial optimization algorithm. Another example, Forster et al. produced a traffic light system-based framework within a national context. Despite its efficiency in expediting decision-making, this framework introduced bias due to its reliance on subjective interpretation. A machine learning model could help address this issue.

Machine learning, a branch of artificial intelligence, enables computers to automatically learn and improve from experience. Large quantities of data are fed into statistical algorithms for pattern identification, enabling them to predict outcomes on unseen data. Compared to traditional experimental methods, ML offers increased speed, scalability, and cost-effectiveness.

Recognizing these advantages, many researchers have selected a ML approach. For example, Zhu et al. predicted biochar yield from pyrolysis in biomass with a ML

algorithm. Another study used ML modeling to rank countries suitability for implementing five greenhouse gas removal technologies (Asibor et al. 2023).

Applying a similar methodology, our study assigned a 'suitability level' to each Virginia county, denoting feasibility of implementation for three CDR technologies.

## II. METHODS

Our methodology was modeled after the Asibor et al. 2023 study, outlined in Figure 1. We specifically focused on three CDR technologies: (1) reforestation, (2) EW, and (3) biochar. We collected several input variables to evaluate the suitability of these technologies, including income, drought index (DSCI), quarterly temperature and precipitation, forest and agricultural land cover, watershed, biomass, power, and water availability.

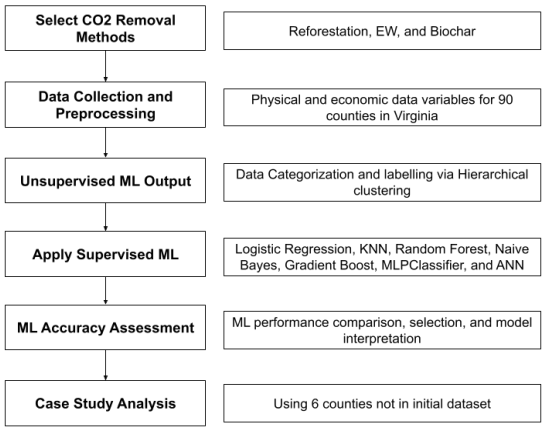


**Figure 1**: Flowchart of methodology.

### *A. Data collection and preprocessing*

The data in this study was scraped from several resources. We considered data from 90 counties in Virginia, excluding cities. This data spans a recent 14-year period (2008-2022). The exclusion of some counties and all cities was due to data collection challenges. Forestation and agricultural land cover percentages were collected from the Multi-Resolution Land Characteristics (MRLC) Consortium. Agricultural land cover was calculated by summing all crop percentages in the MRLC dataset.

Personal income per capita (in dollars) describes the financial capacity of a county, sourced from two resources: Fred, Economic Data and Bureau of Economic Analysis (BEA) Data. The drought index (DSCI) measures drought level, possible values from zero to 500. If zero, none of the area is abnormally dry, or in drought. If 500, all of the area is in D4, exceptional drought. We calculated quarterly mean temperature and precipitation for each county. Our initial attempt at clustering suggested coastal counties as outliers. To better distribute the counties among our four suitability clusters, we added a watershed variable. The watershed predictor indicates if a county is in the Chesapeake Bay Watersheds, referencing a map sourced from the Virginia Department of Conservation and Recreation. This watershed has a tremendous impact on the quality of Virginia's ecosystem, as it covers well over half of Virginia. This raw data was averaged over the past 14 years for each column mentioned above per Virginia county.

The biomass, water availability, and renewable energy product datasets were supplied from the University of Virginia's Environmental Institute. Biomass is measured in dry tonnes. Water availability is the projected available water for the year 2030 after accounting for usage by municipal, industrial, and agricultural sectors, measured in million tons. Energy is the net renewable energy production per county, measured in MWh. Each of these factors contained extreme outliers and zero-inflation. In order to address this, we applied a log function. Data normalization of the entire dataset was achieved by applying 0-1 min-max scalar (*sklearn*) to address the differences in scale and range of the variables. This normalization technique minimizes bias and improves modeling and prediction.

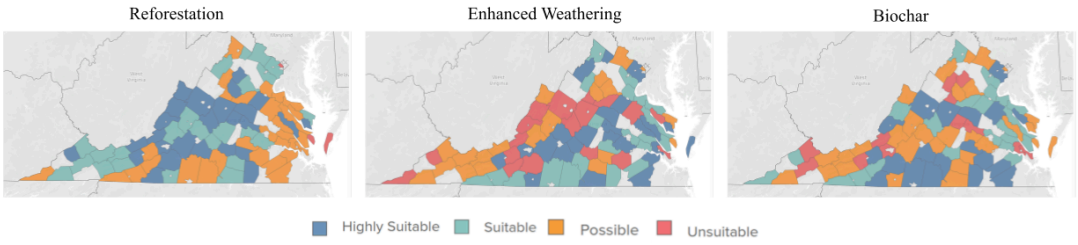


**Figure 2**: Suitability levels of VA counties for each GGR method. Blank spaces represent the the six counties placed in the validation set.

**Table 1:** Major and minor input variables for each GGR method, adapted from Asibor et al. (2023).

| GGR Method | Major Deployment Requirement | Minor Deployment Requirement |
|---|---|---|
| Reforestation | Climatic factor (average quarterly temperature and precipitation*), Forest land use (FLu) | Water Availability |
| Enhanced Weathering (EW) | Low Carbon Energy Available, Income | Average quarterly temperature and precipitation*, Agricultural land use (AgLu) |
| Biochar | Biomass Availability | Average quarterly temperature and precipitation*, Income, AgLu |

* This consists of a total of 8 variables which comprise data for each quarter in a year for temperature (Q1T, Q2T, Q3T, Q4T) and precipitation (Q1P, Q2P, Q3P, Q4P).

Then, we performed a random train/test split using a 80:20 ratio, after removing six validation counties–Accomack, Hanover, Rockingham, Fauquier, Greensville, and Wise. These six counties were specifically chosen due to their even distribution throughout Virginia.

We conducted our modeling in two stages. After preprocessing, we input the above variables into the unsupervised clustering algorithm to derive the output land suitability for each county per CDR technology. For the second stage, we fed the newly labeled dataset into eight supervised ML models and assessed model accuracy.

*B. Hierarchical clustering*

We utilized hierarchical clustering, an unsupervised machine learning method, to group VA counties into specific clusters based on the similarity of input variables. Without specific human-defined cut-off levels, the Agglomerative Clustering library from the sklearn.cluster module created distinct cluster groupings. For each CDR technology, our 84 train/test counties were divided into four clusters to represent

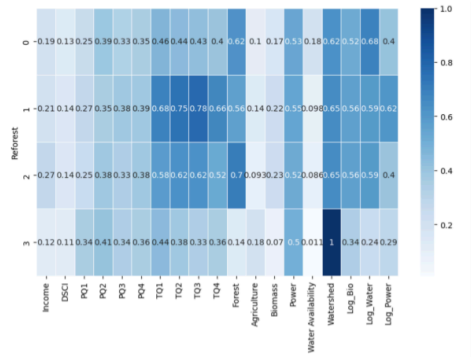


**Figure 3:** Average value for each input variable per cluster for reforestation. Clusters 0, 1, 2, and 3 were assigned the following labels respectively: *suitable, possible, highly suitable, unsuitable*. The darker the blue, the higher the average value.

**Table 2:** Description of suitability levels for VA counties, adapted from Asibor et al. (2023).

| Suitability Category | Description |
|---|---|
| Highly suitable | Satisfies major and minor requirements for deployment |
| Suitable | Satisfies some major and minor requirements for deployment |
| Possible | Satisfies core geo-locational requirements which are not likely to change in a long time (climatic factors and GSP), but currently lacks the highly dynamic requirements such as GNI, LCEA, BA. Given the presence of required geo-locational indicators, the suitability status of nations in this category are expected to improve with improvement in their dynamic indicators. This improvement could be because of factors ranging from economic development to international collaborations |
| Unsuitable | Satisfies neither major and nor minor requirements for deployment |

**Table 3**: Description of supervised ML models, adapted from Asibor et al. (2023).

| Algorithm | Description |
|---|---|
| Logistic regression | A classification algorithm that predicts the likelihood of a dependent variable (usually binary) belonging to a category. |
| Decision Tree | This interpretable algorithm performs by splitting values of data features into branches at decision nodes until a final decision output is established. KNN is a non-parametric algorithm requiring little to no training that makes its selection based on the proximity to other data points regardless of what feature the numerical values represent. |
| K-Nearest neighbor | KNN is a non-parametric algorithm requiring little to no training that makes its selection based on the proximity to other data points regardless of what feature the numerical values represent. |
| Naïve Bayes | This algorithm is based on the Bayesian theorem which updates the prior knowledge of an event with the independent probability of each feature that can affect the event. |
| Random forest | The algorithm is an ensemble of decision trees characterized by improved accuracy. It operates by generating a multitude of decision trees and uses either the modal vote or average prediction for classification or regression tasks, respectively. |
| Gradient boost | This ensemble algorithm combines multiple weak algorithms to obtain an improved output. |
| MLPClassifier | A feed-forward ANN which implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation. |
| Artificial Neural Network | ANN, inspired by the biological neural networks of the human brain, is made up of input, hidden and output layers, as well as a number of parallel-interconnected neurons in each layer. It can be trained to recognise patterns, classify data, and predict future events. |

their suitability level for each of the CDR techniques (Figure 2). The number of clusters was decided by viewing an evidence lower bound (ELBO) plot comparing the number of clusters to the sum of square error.

Using major and minor requirements in Table 1 as a guide, the resulting clusters were then sorted into four levels of suitability: highly suitable, suitable, possible, and unsuitable. Descriptions of these suitability levels can be found in Table 2. Sorting the clusters into the proper suitability classification was based on a comparative assessment of the mean cluster value for each variable. For example, when considering reforestation, the cluster with the highest average percentage of forest land, and the highest water availability and temperature was deemed highly suitable (Figure 3). In general, the higher the mean value for variables of interest, the more suitable the cluster was with the exception of DSCI (which favored small values). For each CDR technology, the 84 counties (train and test) were assigned four separate suitability labels.

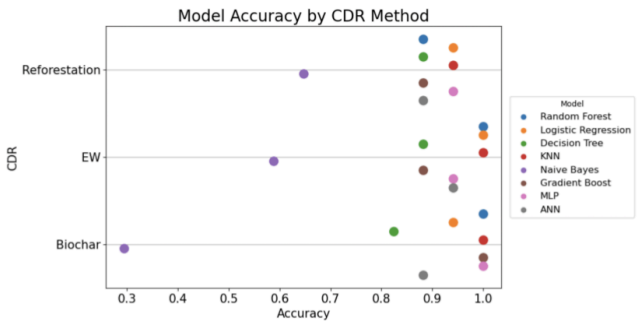


**Figure 4:** Comparison of eight ML algorithms performance per CO2 removal method on testing dataset. The neural network had the highest accuracy for all three GGR methods.

### *C. Supervised ML modeling*

The newly labled dataset was input into eight supervised ML algorithms, detailed in Table 3. Supervised learning models involve training data to understand the relationship between the input variables and the outcome variable, suitability level. Specifically we applied Random Forest, Logistic Regression, Decision Tree, KNN, Naive Bayes, Gradient Boost, MLPClassifier, ANN. Code for each algorithm was written in Python using the respective ML algorithm libraries imported from the *sklearn* module. This process was repeated separately for each CDR technology.

## III. RESULTS

The eight ML algorithms used for predicting suitability level were evaluated on the test set based on their respective performance accuracies (the fraction of correctly classified samples) as shown in Figure 4. Overall, the performance accuracies were very high, ranging from 82.4 - 100% with the exception of the underperforming Naive Bayes model, which had accuracies ranging from 29.4 - 64.7%. The Random Forest, Logistic Regression, KNN, and MLPClassifier algorithms had the highest accuracies overall while Naive Bayes had the lowest. Because machine learning is a fairly opaque process, it is challenging to target a specific reason behind the variation in performance accuracy for each CDR technology. The differences could be due to algorithmic strengths and weaknesses or data characteristics. The performance accuracies were highest on average for enhanced weathering (EW).

The feature importance score represents the impact of each input variable on the output variable. Typically, all importance scores sum to 1. We calculated these scores for

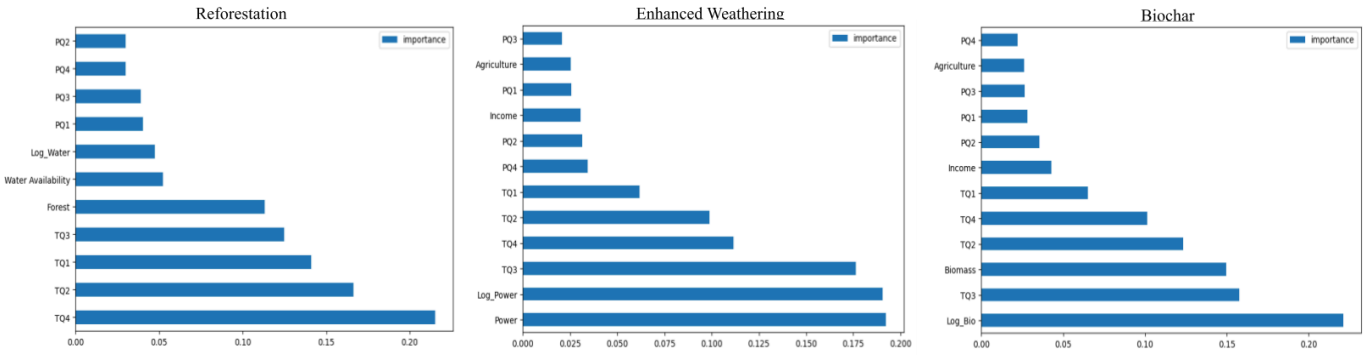


**Figure 5**: Feature importance scores from random forest models for each GGR method. The higher the score, the higher the impact of the input variable on suitability level.

each supervised ML model, but only displayed the importance scores from the Random Forest model (Figure 5).

Feature importance varied across the different removal techniques, but remained relatively consistent across models. Temperature was an overwhelmingly important feature for models for all three removal techniques. Income and agricultural land use were important features for EW and biochar, but not as important for reforestation. Renewable resource energy availability was the most important feature for EW and biomass availability was the most important feature for biochar, as expected from previous research (Asibor et al. 2023). However, forest cover was not as important for reforestation as temperature. Water availability was an important feature for reforestation, but not an important feature for EW and biochar.

### *A. Model Interpretation*

For each CDR technology, we compared the supervised ML model feature importance scores to literature expectations of the relationship between input variables (bio-geophysical and techno economic factors) and suitability levels. For reforestation, climatic factors, such as temperature and precipitation had the highest feature importance scores. Similarly, Asibor et al. (2023) found that climate factors had the highest impact on forestation deployment. For example, tropical regions close to the equator with high quarterly temperatures and high average rainfall are highly suitable for forestation (Favero et al. 2018). The forest land area feature importance score was close behind these climate factors, correlating with literature expectations. Land availability is a major factor in reforestation, behind climate factors (Fuss et al., 2018).

Low-carbon emission energy availability had the highest importance score for EW, followed by climate factors, then income and agriculture land area. In concurrence, Smith et al. states the most important factor for EW is low-carbon emission energy availability because rock splitting is essential for small particulate production. However, the ML model overestimated the importance of climate factors, while simultaneously underestimating income (Strefler et al. 2018). For biochar, biomass availability importance highly correlates with literature expectations. Biomass availability is the most important factor for biochar, as biomass is the base ingredient for biochar (Smith 2016). The ML model overestimated the importance of temperature. Asibor et al. does not list temperature as a major factor in biochar implementation.

### *B. Modeling case study*

We assessed the model performance accuracy on the six withheld counties–Accomack, Hanover, Rockingham, Fauquier, Greensville, and Wise. This was our validation dataset. We ran each supervised ML model on this subset to predict their suitability levels. The majority of the models predicted the same suitability levels for the 6 counties, across all CDR technology.

#### *i. Reforestation*

Figure 6 displays the suitability levels of each CDR technology by county. Table 1 explains that reforestation largely depends on (1) precipitation, (2) temperature, (3) forest land use, and (4) water availability. Forestry is Virginia's third leading industry and Brunswick county leads the state in lumber production. Out of the 6 validation counties, our model characterized only Hanover as *suitable* for reforestation. Surprisingly, many of Hanover's

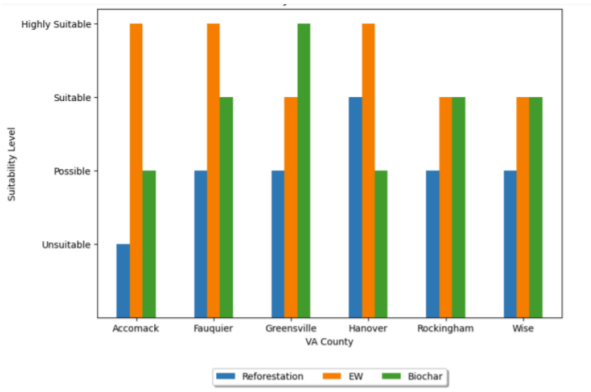


**Figure 6**: Predicted suitability levels of the six counties in our validation set, separated by GGR method. Reforestation is in blue, EW in orange, and biochar in green.

neighboring counties were labeled *possible* for reforestation by the unsupervised ML clustering. It is important to note that Hanover borders Richmond, a major city in Virginia. In relation to the other six counties, Hanover has the fourth greatest forest land area percentage (Rockingham had the highest forestation). In Figure 2, the highly suitable counties for reforestation tended to be in western Virginia, contrary to our validation results.

Hanover announced a new solar policy in 2023. To make room for the new solar farms, many forest patches had to be cleared consequently. However, Hanover claims to have plans for replanting these patches in the policy agreement.

Accomack county was categorized as unsuitable for reforestation. This is logical because of its coastal location and small forest land use percentage. Also, Accomack has the highest precipitation and lowest water availability values, compared to all other counties in the validation set.

*ii. Enhanced Weathering*

The suitability level for EW is largely reliant on (1) low carbon energy availability, (2) income, (3) temperature, (4) precipitation, and (5) agricultural land use. Accomack, Fauquier, and Hanover counties were categorized as *highly suitable* for EW. Most notably, these three counties all had high power values, and high agricultural land area percentages. The other three counties: Rockingham, Wise, and Greensville were all considered *suitable* for EW by our model. Even though Greensville possessed the highest power value, it was not labeled *highly suitable*. Rockingham had the lowest power value of the six counties. Geographically, the three most eastern counties were categorized as highly suitable, correlating with Figure 2.

Interestingly, EW is highly beneficial in an agricultural landscape because the technique can improve crop yields. Accomack, Fauquier, and Hanover all have high agricultural land use percentages. Coastal counties are more exposed to rapid weathering, thus accelerating the EW process. The Virginia coastal plain contains lots of eroded clay, sand, and gravel which is ideal for enhanced weathering.

*iii. Biochar*

Virginia has a strong initiative for biochar implementation because of the state's overall land suitability. As mentioned in Table 1, biochar is highly dependent on (1) biomass availability, (2) temperature, (3) precipitation, and (4) income. The model categorized Greensville as *highly suitable* for this CDR technology. Greensville had the highest biomass availability possible and the highest quarterly temperatures, compared to the other six counties. Geographically, Greensville is located in southeast Virginia. Counties on the southern border of Virginia tended to be categorized as *suitable* or *highly suitable* by our model, correlating with previous literature findings. It makes sense that biochar is most prominent in Southern Virginia because the Roanoke River basin region provides logging excess for biomass energy. Additionally, there are power plants near Greensville, like in Southampton, that rely exclusively on wood fuel.

On the other hand, Accomack and Hanover counties were categorized as *possible* for biochar. These eastern Virginia counties had high agricultural land use, especially in Accomack. Most notably, Accomack and Hanover had the lowest temperatures in all quarters.

*C. Limitations*

The methodology was constrained due to several limitations occurring in the data collection process. We collected data on VA counties, excluding cities. This exclusion occurred due to the inconsistencies in data availability for cities, particularly cities surrounded by counties present in the dataset. Furthermore, the substantial proportion of developed land in cities renders them inefficient for the implementation of removal techniques

(Fuss et al. 2018). The accessibility of data constrained the consideration of direct air carbon capture and storage (DACCS) and bioenergy with carbon capture and storage (BECCS). Given additional resources, further research could explore suitability prediction for both cities and DACCS and BECCS removal techniques

Additionally, due to the unsupervised nature of the clustering technique used to categorize each county based on suitability, we cannot be fully confident in our categorizations. The lack of total confidence in the generated suitability variable limits our confidence in our supervised model predictions. Finally, we were interested in exploring the relationship between Virginia county suitability levels for each CDR method and their environmental policies. Specifically we aimed to explore land use ordinances, laws, and conservation easements to more deeply understand and validate our model results. Unfortunately, we were limited by a combination of lack of data availability and difficulty in the data collection process which exceeded our timeframe.

## IV. CONCLUSION

Machine learning approaches play an integral role in predicting the suitability of Virginia counties for three greenhouse gas removal technologies, (1) reforestation, (2) EW, and (3) biochar. Identifying suitable locations enables more efficient implementation of these costly technologies, ultimately advancing greenhouse gas removal efforts. Our dataset consisted of environmental and economic data collected from publicly available resources. Following the methodology proposed by Asibor et al., we employed both supervised and unsupervised methods. Our unsupervised clustering algorithm assigned labels for the 84 counties in the training set. Eight supervised models then predicted these labels. Accuracy ranged from 82.4-100% for all models excluding Naive Bayes, which had the lowest accuracy. The Random Forest, Logistic Regression, KNN, and MLPClassifier models exhibited the highest accuracies across removal techniques, with the EW technique having the highest overall accuracy. We assessed model performance on validation dataset, revealing EW to be highly suitable in Accomack, Fauquier, and Hanover. Cities were excluded from the models due to inconsistent data availability and removal techniques were restricted due to accessibility of data. The utilization of ML enhances the implementation of greenhouse gas removal techniques, proving beneficial in driving Virginia towards a reduction in CO2 emissions.

## V. DATA

The code/data of the project can be found at: https://github.com/hrnur/uvaeicapstone

## VI. REFERENCES


[1] Akyuz, F. A. (2017). Drought Severity and Coverage Index. United States Drought Monitor. https://droughtmonitor.unl.edu/About/AbouttheData/DSCI.aspx

[2] Asibor, J. O., Clough, P. T., Nabavi, S. A., & Manovic, V. (2023). A machine learning approach for resource mapping analysis of greenhouse gas removal technologies. Energy and Climate Change, 4, 100112. https://doi.org/10.1016/j.egycc.2023.100112

[3] Biomass Energy in Virginia. (n.d.). http://www.virginiaplaces.org/energy/biomass.html

[4] Bureau of Economic Analysis, U.S. Department of Commerce. Personal Income Per Capita. https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas.

[5] ClimateSeed (2024). Carbon Dioxide Removal (CDR) Methods. https://climateseed.com/blog/carbon-removal-methods

[6] Favero, A., Sohngen, B., Huang, Y., & Jin, Y. (2018). Global cost estimates of forest climate mitigation with albedo: A new integrative policy approach. Environmental Research Letters, 13(12), 125002. https://doi.org/10.1088/1748-9326/aaeaa2

[7] Fred Economic Data, St. Louis Fed. Per Capita Personal Income by County, Annual: Virginia. https://fred.stlouisfed.org/release/tables?eid=268980&rid=175.

[8] Fuss, S., et al. (2018). Negative emissions—Part 2: Costs, potentials and side effects. Environmental Research Letters, 13(6), 063002. https://doi.org/10.1088/1748-9326/aabf9f

[9] IPCC (2022). Carbon Dioxide Removal. IPCC AR6 WGIII Factsheet CDR. https://www.ipcc.ch/report/ar6/wg3/downloads/outreach/IPCC_AR6_WGIII_Factsheet_CDR.pdf.

[10] Lashof, D. A., & Ahuja, D. R. (1990). Relative contributions of greenhouse gas emissions to global warming. Nature, 344(6266), 529–531. https://doi.org/10.1038/344529a0

[11] NOAA National Centers for Environmental Information (2024). Climate at a Glance: County Time Series. https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/time-series.

[12] *Rocks of the Coastal Plain*. Earth@Home. (2023, September 20). https://earthathome.org/hoe/se/rocks-cp/

[13] Schuiling, R. D., & Krijgsman, P. (2006). Enhanced Weathering: An Effective and Cheap Tool to Sequester Co2. Climatic Change, 74(1–3), 349–354. https://doi.org/10.1007/s10584-005-3485-y

[14] Smith P. (2016). Soil carbon sequestration and biochar as negative emission technologies. Glob Chang Biol. 22(3):1315-24. doi: 10.1111/gcb.13178.
[15] Smith, P., et al. (2016). Biophysical and economic limits to negative CO2 emissions. Nature Climate Change, 6(1), 42–50. https://doi.org/10.1038/nclimate2870
[16] *Solar policy: Hanover County, VA*. Solar Policy | Hanover County, VA. (n.d.). https://www.hanovercounty.gov/1301/Solar-Policy
[17] Strefler, J., Amann, T., Bauer, N., Kriegler, E., & Hartmann, J. (2018). Potential and costs of carbon dioxide removal by enhanced weathering of rocks. Environmental Research Letters, 13(3), 034010. https://doi.org/10.1088/1748-9326/aaa9c4
[18] Sweeney, C., et al. (2015). Seasonal climatology of CO 2 across North America from aircraft measurements in the NOAA/ESRL Global Greenhouse Gas Reference Network. Journal of Geophysical Research: Atmospheres, 120(10), 5155–5190. https://doi.org/10.1002/2014JD022591
[19] United States Department of Agriculture (2023). Cropland Data Layer. https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php
[20] Virginia Department of Conservation and Recreation. Virginia's Major Watersheds. https://www.dcr.virginia.gov/soil-and-water/wsheds.
[21] Virginia Department of Social Services (2024). Cities/Towns/Counties with Region and FIPS. https://www.dss.virginia.gov/family/cc_providertrain/region_lookup.pdf.
[22]Wang, J., & Wang, S. (2019). Preparation, modification and environmental application of biochar: A review. Journal of Cleaner Production, 227, 1002–1022. https://doi.org/10.1016/j.jclepro.2019.04.282