Apple cultivars and climate change in the Eastern Mountain region of the United States

Abhay Baliga, Jayme Fisher, Willie Man

Introduction

Extensive research exists regarding the effects of climate change on agriculture. However, there is often a gap between researchers and decision makers, particularly in industries with limited resources. The apple industry is one such example. Though apples are the most consumed fruit in the United States, apple orchards occupy less than 1% of the 880 million acres of United States farmland (Industry at a Glance, n.d.; National Agricultural Statistics Service, 2024). To bridge this gap, we have developed a proof of concept for an apple manufacturer, providing insight into the resilience of cultivars in the Eastern Mountain region of the United States. Our findings suggest that this region will shift into the next USDA Plant Hardiness Zone within ten years, and that two apple cultivars—Cortland and Franklin Cider—may no longer be viable. If valuable, these insights could pave the way for other regions and crops, expanding both the accessibility and relevance of essential climate data.

Background

Industry. Climate change is one of a number of risks in the apple industry, alongside production, finance, and marketing (Morton et al., 2017). From a production standpoint, the industry is stagnant, having generated roughly 10 billion pounds per year for the past twenty years (Weber et al., 2023). The number of orchards has declined from 21,084 in 2002 to 20,530 in 2022 (National Agricultural Statistics Service, 2024). In addition, while prices have increased, profit margins remain slim, particularly for manufacturers.

The relationship between climate change and apple production is surprisingly complex. “Changes in climate interact with other environmental and societal factors in ways that can either moderate or intensify its impacts on production systems” (Morton et al., 2017). For example, while higher levels of carbon dioxide generally increase yields, research suggests that when combined with heat stress, drought, and nutrient deficiencies, carbon dioxide inhibits growth. While research into the effects of climate change is ongoing, growers and scientists have already begun developing and adopting new strategies. “In Washington state, scientists have developed a spray made out of cellulose from wood pulp to help insulate fruit. In Georgia, some farmers are growing citrus, or olives” (Hoplamazian, 2023).

Growers also use cultivars to mitigate risk. Cultivars, or cultivated varieties, are types of apples specifically bred for desired traits such as color, size, and taste. There are dozens of existing cultivars, including popular varieties such as Golden Delicious, Red Delicious, Rome, and York, and “the apple cultivar situation is constantly changing” (Peter, 2024). A number of factors influence cultivar selection. Temperature is one of the most pivotal. Apples require cold winters to remain dormant. Warmer winters can lead to premature blossoming, increasing the risk of crop loss due to unexpected frosts. As temperatures fluctuate, the risk escalates, prompting some growers to experiment with new cultivars that are better adapted to changing conditions.

Data. Agriculture in the United States benefits from a wealth of information, published by both national and state governments. This data encompasses everything from agricultural guides to climate forecasts and weather reports. Though extensive, effectively leveraging this vast array of data often requires a nuanced understanding of data retrieval and data storage. To address this need, data consultancies like FixCarbon play a crucial role. FixCarbon, a sponsor of this project, specializes in collecting, transforming, and analyzing

geospatial data for permanent agriculture supply chains. The company has access not only to government datasets but also data from third-party brokers such as Oikolab.

Manufacturers generally recognize the value of climate data. Growers, on the other hand, tend to be more skeptical. Though some acknowledge trends in crop yields, the perception tends to be that each year is a new year in the fruit business, and that data is no substitute for experience. Thus manufacturers, rather than growers, are the natural audience for agricultural data products.

Client. Our client for this proof-of-concept is a regional manufacturer of apple products based in the Eastern Mountain region of the United States. Collaborating closely with local growers, this manufacturer uses apples that do not meet retail standards to produce items like apple butter, apple juice, and applesauce. Importantly, the manufacturer uses a blend of cultivars in its product lineup. As climate change begins to affect availability, the manufacturer will need to work with growers to identify new cultivars and reformulate its products.

Approach

Data Source. Our proof-of-concept is based on data from four distinct sources: the United States Department of Agriculture (USDA), the National Aeronautics and Space Administration (NASA), Oikolab, and the Stark Bro's website.

CroplandCROS is an application developed by the National Agricultural Statistics Service (NASS) within the USDA. It uses satellite imagery to locate specific crops within the continental United States. This tool provides data with a spatial resolution of 30 meters that is available for download as a GeoTIFF file. For this project, we used the Cropland Data Layer (CDL) for 2023, filtered for apples in the Eastern Mountain region. We used this data to define the region for our regression analysis.

NEX-GDDP-CMIP6 is a dataset comprised of downscaled climate model and scenario projections. It was developed by NASA in 2022 in collaboration with governments and researchers around the world. Variables such as precipitation, radiation, and temperature are available from 1950 through 2100. The data is stored in an Amazon S3 Bucket as an Xarray Dataset with a spatial resolution of 25 kilometers. For this project, we retrieved minimum temperature from 2015 to 2050 for all available models and scenarios in the Eastern Mountain region. This data served as the features in our regression analysis.

There are four climate scenarios in the NEX-GDDP-CMIP6 dataset, also known as Shared Socioeconomic Pathways (SSPs). These SSPs represent specific greenhouse gas emission scenarios. For example, in the worst-case scenario (SSP585), no climate measures are taken, fossil fuel consumption remains unchanged, and radiative forcing reaches 8.5 watts per square meter. Conversely, in the best-case scenario (SSP126), climate measures are taken, and radiative forcing is limited to 2.6 watts per square meter. These scenarios serve as foundational inputs for the NEX-GDDP-CMIP6 climate models, which simulate a range of potential outcomes vis-à-vis temperature, precipitation, and other such variables.

Oikolab provides access to weather datasets through an application programming interface (API). For this project, we used the ERA5 dataset, which provides historical reanalysis from 1940 through the present with a spatial resolution of 28 kilometers. FixCarbon provided an API wrapper that accepts a Shapely Polygon and returns netCDF dataset, which is compatible with Xarray. ERA5 served as the ground truth in our regression analysis.

Stark Bro's is a nursery offering more than 70 apple cultivars. On the Stark Bro's website, each cultivar is described in detail with characteristics such as bloom time, chill hours, and years to bear fruit. For this project, we scraped the USDA Plant Hardiness Zone associated with each cultivar. This allows us to assess the resilience of specific cultivars as climate conditions change. USDA Plant Hardiness Zones are based on 30-year average temperature minimums (Figure 1). Each zone represents 10 degrees fahrenheit, and each half zone represents 5 degrees. There are 13 zones and 26 half zones in total across the United States.

Data Selection. The entire NEX-GDDP-CMIP6 dataset is 30 terabytes. To obtain a more manageable dataset, we used the CroplandCROS data to define and select a subset (Figure 2). This involved using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to assign orchards to clusters. HDBSCAN was selected over other clustering algorithms because of its handling of outliers, as our focus was on areas densely populated with orchards, rather than all orchards in the Eastern Mountain region. After clustering, the orchard clusters were transformed into polygons using Alpha Shape. These polygons were then saved with GeoPandas and used to clip the NEX-GDDP-CMIP6 and ERA5 datasets.

Data Preparation. The spatial resolution of the NEX-GDDP-CMIP6 dataset is 25 kilometers whereas the ERA5 dataset has a resolution of 28 kilometers. This means that the coordinates of the two datasets do not align. To merge the datasets, we applied linear interpolation, converting the resolution of NEX-GDDP-CMIP6 from 25 to 28 kilometers. The data was then segmented into four distinct subsets: train, validate, test, and project. Finally, to optimize storage and enhance processing efficiency, these datasets were compressed and saved in the Parquet format, which is particularly effective for handling large datasets.

Because our data is a time series, the data in the train, validate, test, and project datasets cannot be random. Instead, we used time periods.

- Train: 01/01/2015 - 12/31/2021
- Validate: 01/01/2022 - 12/31/2022
- Test: 01/01/2023 - 12/31/2023
- Project: 01/01/2024 - 12/31/2050

Figure 1. USDA Plant Hardiness Zones published by the Agricultural Research Service.

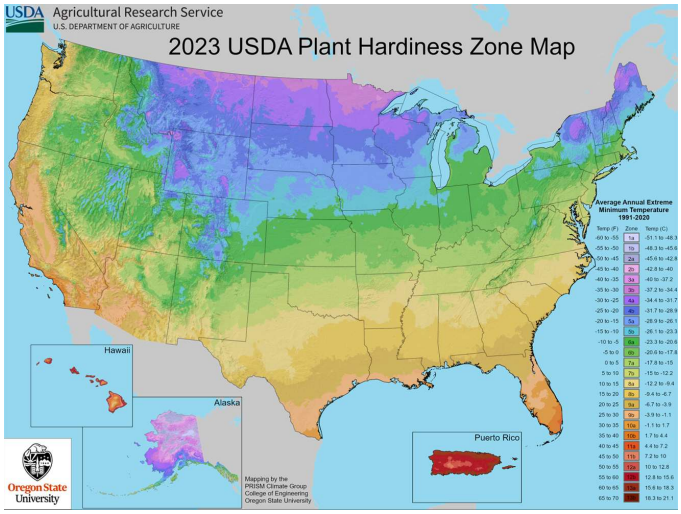
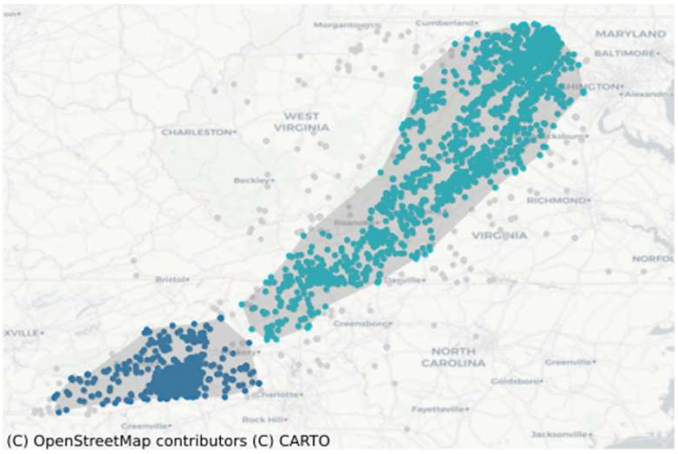


Figure 2: Orchards and orchard clusters in the Eastern Mountain region of the United States.



Regression. Remember that nurseries rely on USDA Plant Hardiness Zones to determine where apple cultivars can survive and thrive, and that USDA Plant Hardiness Zones are based on temperature. In this context, temperature is an indicator of climate resilience. Thus, we trained an elastic net model to predict minimum temperature. This approach ultimately allows us to update the USDA Plant Hardiness Zones and identify cultivars that are no longer viable.

Chase Dwelle at FixCarbon recommended using an elastic net model. Elastic net both limits overfitting with L1 regularization and performs feature selection with L2 regularization. We used all 80 model-scenario combinations in the NEX-GDDP-CMIP6 dataset as features, along with latitude and longitude. We used minimum temperature from the ERA5 dataset as our target variable. Note that the model assumes each coordinate is independent.

As a baseline for comparison, we averaged the 80 model-scenario combinations in NEX-GDDP-CMIP6 and calculated the coefficient of determination (R^2) and root mean square error (RMSE). We also trained an ordinary least squares regression model using the same features and target.

In our modeling efforts, we considered using both daily data and rolling data. We ultimately determined that a 30-day rolling window delivered the best results. Daily data is noisy, complicating pattern recognition. In contrast, rolling data is less erratic, facilitating effective learning (Figure 3). In addition, several climate models were missing values during leap years. To address this, we applied forward fill.

Evaluation. Our final regression model maximizes R^2 and minimizes RMSE. Results are available in Table 1 and Table 2.

Figure 3. Minimum average temperature in the Eastern Mountain region in degrees Fahrenheit for rolling windows of 1, 7, 30, 90, and 180 days.

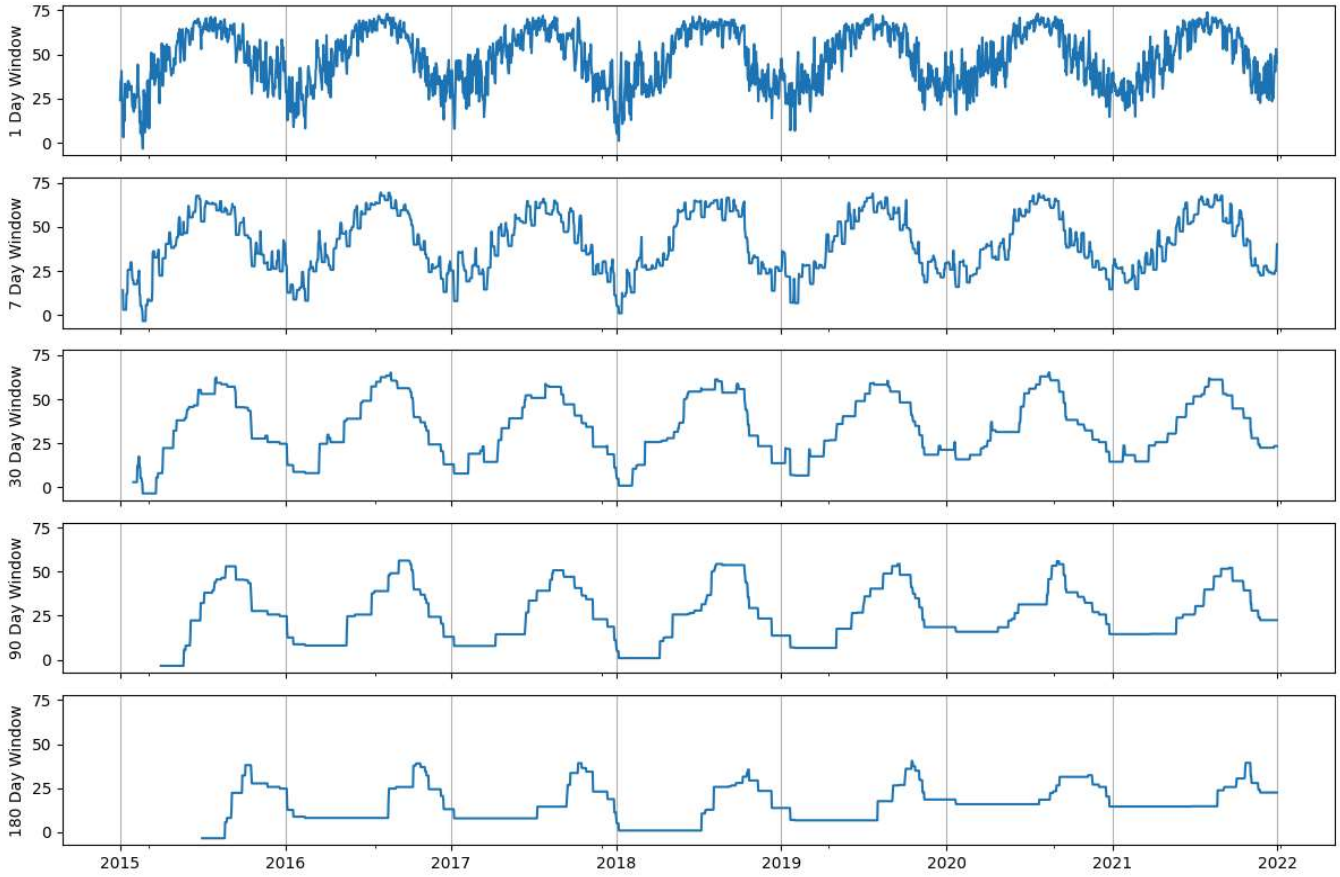


Table 1. Coefficient of determination (R^2) for the baseline, the ordinary least squares, and elastic net models.

	Train Data	Validate Data	Test Data
Baseline Model	0.76	0.84	0.75
Ordinary Least Squares Model	0.94	0.88	-
Elastic Net Model	0.92	0.89	0.89

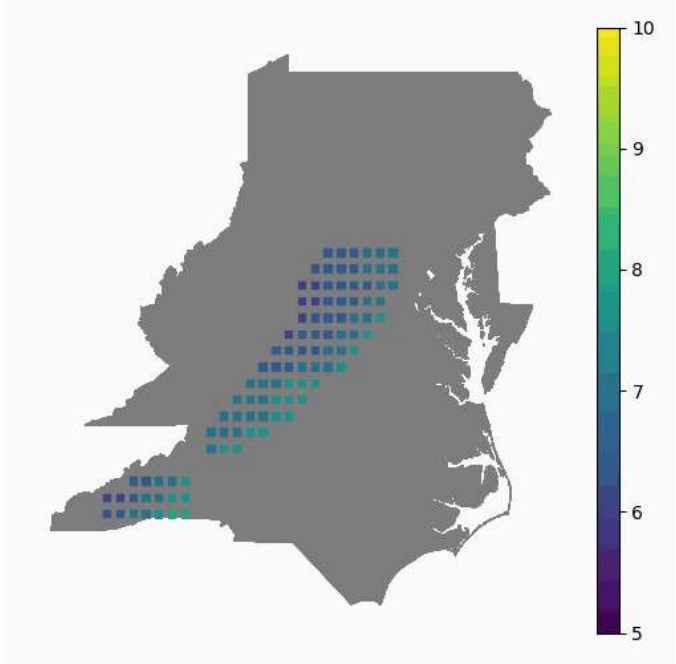
Table 2. Root mean squared error (RMSE) for the baseline, ordinary least squares, and elastic net models.

	Train Data	Validate Data	Test Data
Baseline Model	8.57	7.57	8.34
Ordinary Least Squares Model	4.33	6.50	-
Elastic Net Model	5.12	6.35	5.67

Results

Our predictions suggest that while minimum temperatures increase up to 12 degrees Fahrenheit by 2050, the change in USDA Plant Hardiness Zones is more gradual, due in large part to the 30-year rolling average in the calculation. Figure 4 illustrates how these zones change over time.

Figure 4: USDA Plant Hardiness Zones in the Eastern Mountain region from 2014 to 2050 based on elastic net regression.



Assuming the USDA continues to use a 30-year rolling average for risk assessment, most apple cultivars will continue to be viable. The cultivars in our research can succeed in USDA Plant Hardiness Zones 2A to 9B. In the Eastern Mountain region, our predictions range from 5A to 8A, within the approved range for most cultivars. However, two cultivars, the Cortland Apple and the Franklin Cider Apple, may no longer be viable past 2025, when the entire region shifts into 6A. Furthermore, as seen in Figure 5, the majority of the region will be in 7A just after 2030, jeopardizing an additional 13 cultivars, including Ben Davis, Empire, Enterprise, Liberty, and trademarked cultivars CandyCrisp, Hart's Fancy, KinderKrisp, Pioneer Mac, Royal Empire, Scarlet Crush, Red Romance, Ruby Darling, and SnowSweet.

Our elastic net model is robust to regression methods and parameters. However, it is sensitive to decisions about methods of normalization. A rolling mean approach reduced volatility improved our ability to predict trends in the data, even obtaining an R^2 score of 0.98. However, this approach did not allow us to capture actual temperature minimums. Because USDA Plant Hardiness Zones are calculated based on minimums, these values were crucial for trustworthy predictions. A rolling minimum with a shortened window results in an R^2 score of 0.89. This approach better captures our data of interest.

Figure 5: Count of latitude, longitude pairs in each USDA Plant Hardiness Zone. Notice the consolidation over time as temperatures continue to increase.

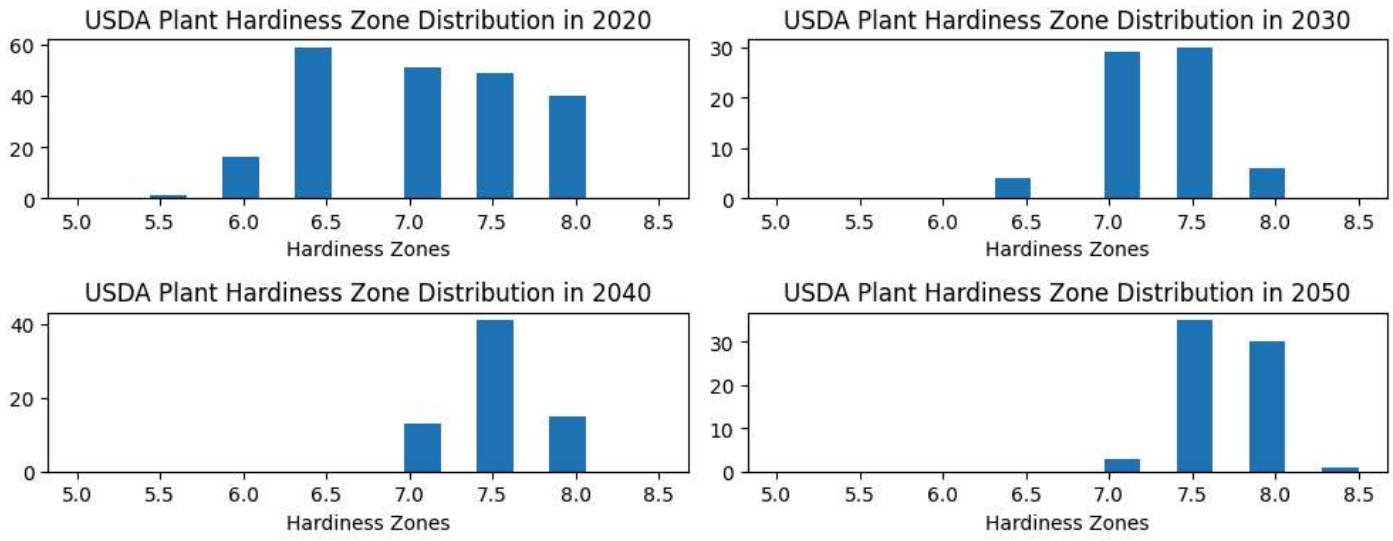
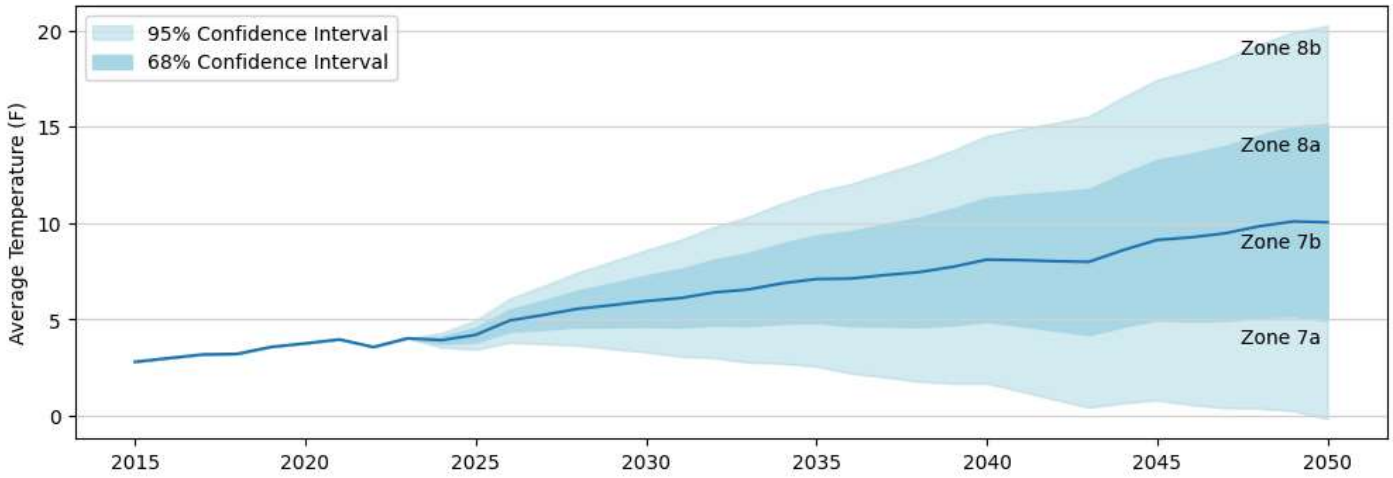


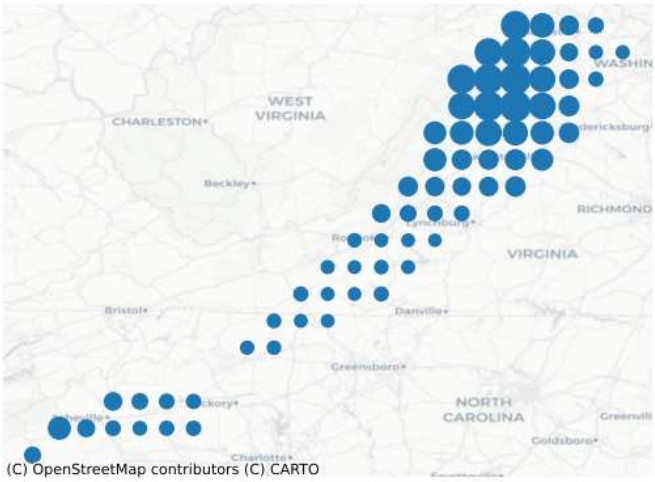
Figure 6: Minimum average temperature in the Eastern Mountain region of the United States with a 30-year rolling window, projected through 2050.



As time progresses, our predictions become less certain. Climate change is complex and trends are difficult to predict. However, it is more likely than not that the region will shift into the next USDA Plant Hardiness Zone within ten years (Figure 6).

It is worth noting that variability exists at the coordinate level (Figure 7). Mountainous regions are particularly difficult to predict, with a maximum RMSE of 9.37. It is possible that mountain temperatures are more variable than temperatures in other topographies. If so, an elevation feature might improve model performance.

Figure 7: Root mean squared error based on latitude and longitude. Larger circles indicate higher levels of uncertainty. The most variable coordinates appear over the Appalachian Mountains on the border between Virginia and West Virginia.



Discussion

Temperature is less relevant in tropical regions where factors such as humidity are more likely to drive resilience. However, retrieving data from climate models and comparing the results to weather trends remains a viable approach. Humidity, precipitation, radiation, temperature, and wind speed are all available globally in both climate and weather data. This suggests that with a similar machine learning pipeline, it should be possible to predict the impact of any of these variables in any region.

Considering temperature specifically, it is worth noting that USDA Plant Hardiness Zones are potentially misleading. Imagine a farmer planning to plant a new orchard. Orchards are generally profitable for 10 to 15 years. Thus, in addition to the official USDA Plant Hardiness Zone Map, the farmer might consult a map with projections for 2040. Because USDA Plant Hardiness Zones are based on 30-year averages, the 2040 projections still incorporate data from 2010. If climate change continues to accelerate, the minimum temperature in 2010 may no longer be relevant. This can be mitigated by reducing the rolling window, though whether growers are willing to consider different metrics is up for debate.

There is another shortcoming in our approach. While our model suggests certain apple cultivars are likely to be resilient based on temperature, it is missing other factors that affect cultivar selection. This includes physical factors such as soil composition and susceptibility to diseases, as well as financial factors such as consumer demand and margins. For example, fresh apples garner significantly higher returns than processed apples (Weber et al., 2023). Thus, there is an incentive to produce the cultivars sold in grocery stores. Incorporating these additional factors would undoubtedly prove valuable for growers and manufacturers alike.

Conclusion

USDA Plant Hardiness Zones, frequently used for crop selection in the United States, are inherently retrospective, based as they are on 30-year averages. Our approach provides a forward-looking alternative, predicting changes to the USDA Plant Hardiness Zone Map. This allows growers and manufacturers of permanent crops to make more informed decisions, even with changing climate conditions.

Acknowledgements

We would like to thank Chase Dwelle from FixCarbon for providing access to NEX-GDDP-CMIP climate data and Oikolab weather data, as well as feedback and support from the earliest stages of problem formation through solution design.

We would also like to thank our contacts at two different manufacturers for providing invaluable insight into climate change and its effects on the agricultural industry.

Statement of Work

Abhay Baliga recreated the function to derive hardiness zones from daily minimum temperatures. He analyzed the results of the model and mapped the cultivars. He developed and finalized several of the visuals. He is also responsible for drafting the results and discussion sections.

Jayme Fisher completed the introduction and background sections of the report. She developed the code defining the Eastern Mountain region; the code combining the climate data and weather data; and the code training an elastic-net model to predict minimum rolling temperature. She also conducted qualitative interviews with individuals at two different manufacturers leveraging agricultural products.

Willie Man scraped data from the Stark Bro's website; retrieved climate data from the NEX-GDDP-CMIP6 Amazon S3 Bucket, retrieved weather data from the Oikalab API; and provided visualizations, including the time horizon plots. He also completed an initial draft of the approach section of the report.

References

Hoplamazian, M. (2023, October 20). A bad apple season has some U.S. fruit growers planning for life in a warmer world. NPR. Retrieved April 17, 2024, from https://www.npr.org/2023/10/20/1207202139/a-bad-apple-season-has-some-u-s-fruit-growers-planning-for-life-in-a-warmer-worl

Industry at a glance. (n.d.). USApple. Retrieved April 17, 2024, from https://usapple.org/industry-at-a-glance

N. Mohammed, I. (2024, January 13). Getting started with NEXGDDP-CMIP6 data. NASAaccess. Retrieved April 17, 2024, from https://imohamme.github.io/NASAaccess/articles/NEXGDDP-CMIP6.html

Morton, L. W., Cooley, D., Clements, J., & Gleason, M. (2017). Climate, weather and apples. Department of Sociology, Iowa State University. Retrieved April 17, 2024, from https://www.climatehubs.usda.gov/sites/default/files/Climate,%20Weather%20and%20Apples_0.pdf

National Agricultural Statistics Service. (2024). Census of agriculture [Dataset]. United States Department of Agriculture. https://quickstats.nass.usda.gov/

Peter, K. (2024). Penn State Tree Fruit Production Guide. The Pennsylvania State University.

United States Department of Agriculture. (2023). 2023 USDA Plant Hardiness Zone Map. https://planthardiness.ars.usda.gov/system/files/National_Map_HZ_8x11_HS_300.png

Weber, C., J. Wechsler, S., & Wakefield, H. (2023). Fruit and Tree Nuts Yearbook [Dataset]. United States Department of Agriculture. https://www.ers.usda.gov/data-products/fruit-and-tree-nuts-data/fruit-and-tree-nuts-yearbook-tables/

Appendix

The code and much of the data for this project is available on GitHub at https://github.com/FixCarbon/um-mads.

Appendix Figure 1. Rolling minimum versus rolling mean in the Eastern Mountain region through 2023. Notice that the rolling minimum in the first figure yields temperatures that are 10 to 20 degrees colder than the rolling mean in the second figure.