CLIMATE ALPHA

Climate Alpha

White Paper (V2.5)

An overview of Climate Alpha’s data engineering, machine learning, and analysis methods

Updated as of 08 July 2022

CONTENTS

1. Introduction 3
2. Product Design Summary 3
3. Pipeline components 4
4. Baseline Forecaster Details 8
5. Scenario Forecaster Details 13
6. Feature Weight Explorer 20
7. Resilience Index ™ 21
8. Climate Price ™ 27

Appendix 28

1. INTRODUCTION

Climate Alpha is an AI-powered analytics platform that drives data-driven real estate strategies. Our models capture the impact of a wide range of variables on land value and property prices under multiple climate change scenarios. We run previously fragmented datasets through machine learning and mixed-modeling techniques to quantify the impact of key variables on land and real estate asset prices.

Powering Climate Alpha is a patent-pending machine learning pipeline that forecasts trends based on customizable scenarios provided by the user in real-time. Climate Alpha's algorithm is currently able to predict the land value in backtesting to within 4% of the actual price for ~38% of US counties, and within 6% for nearly 60% of US counties.

New datasets are continuously added to improve the reliability and explainability of our forecasts. Actual market results are also regularly fed into the models to improve accuracy.

2. PRODUCT DESIGN SUMMARY

Climate Alpha's architecture can be broken down into offline processes which handle the baseline forecasts and online processes which generate real-time scenario-based predictions.

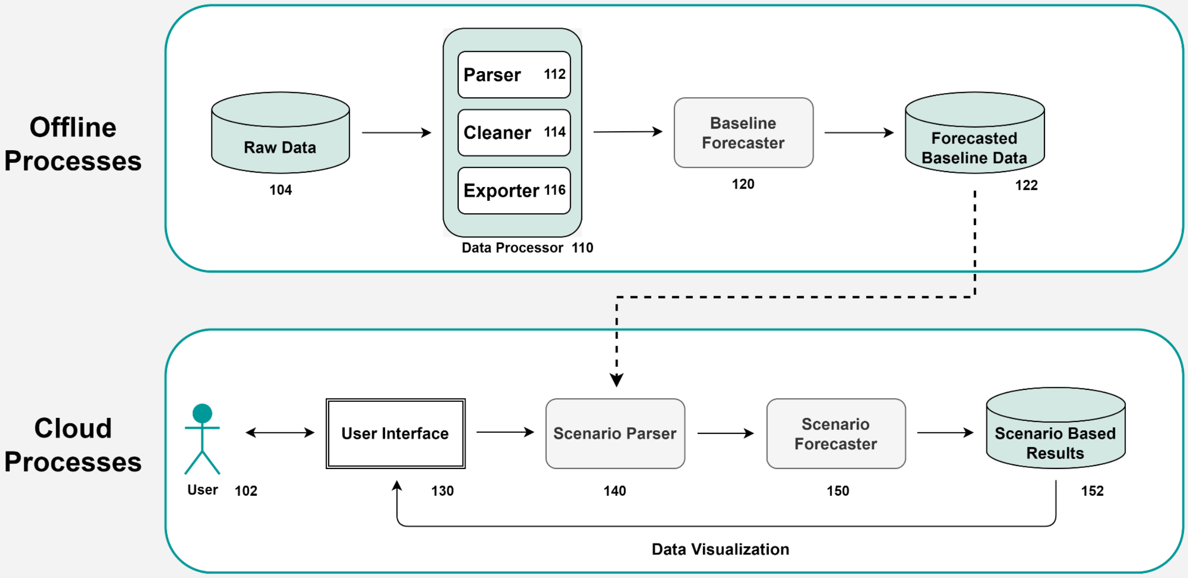


Figure 1

3. PIPELINE COMPONENTS

Overview

The following table summarizes specific ML problems and their solutions in each step of the architecture shown in Figure 1.

Components	Problem Type	Data	Task	Packages and Techniques Used
Data Processor	Data engineering	Raw data from various sources. Types include Time-series, tabular, and GIS data.	Clean, process and merge raw data into a single dataframe for Baseline Forecaster.	- Interpolation - GIS operations - Data Engineering
Baseline Forecaster	Hierarchical time-series forecasting	Grouped time-series data from 2012 to 2019 for approximately 1000 groups (counties).	Forecast trends from processed time-series data.	- LightGBM - ARIMA - Dense Neural Network (NN) - Convolutional NN - Long Short Term Memory (LSTM) NN - Autoregressive NN
Scenario Forecaster	Regression	The forecasted baseline for 1000 counties; each group's data ranging from 2012 to 2040.	Combines forecasted time series and other tabular data to predict the effect of custom scenarios on the target.	- XGBOOST - LightGBM - KNN - Logistic Regression
Scenario Parser	Data engineering	User input in the form of JSON with factors for each feature.	Modified the baseline data with user input factors.	- Data Engineering
Scenario Modulator	Statistical modeling	User input in the form of JSON with factors for each feature.	Modulate predicted results with statistical models to account for shifts in fundamentals.	- Statistics-driven coefficient computation for different sectors (e.g. clean energy adoption and climate change development pathways)

Data Processor

The Data Processor takes raw data from different sources and prepares them for downstream training. For each dataset, we:

a. Interpolate and pad missing time-series data[1]
b. Remove remaining NA values
c. Downscale[2] CBSA/State data into their constituent counties, appending the same data across all the constituent counties
d. Merge all time-series feature data into one data frame for the Baseline Forecaster
e. Merge all non-time series feature data in one data frame for the Scenario Forecaster

Data is collected at the state, county, or CBSA levels and stored in their respective directories. Within each directory, there are subfolders containing the raw data, a python script to clean and merge the raw data, and a final .csv file containing the cleaned dataset.

Initial processing involves data cleaning and renaming specific features for consistency when all the datasets are merged (for instance, all-time series cleaned files at the county level must include a 'County Code' and 'Year,' and should follow the same naming convention). A script is then used to merge cleaned data files. The script is designed to automatically generate merged datasets individually for each geographical level if needed, and a combined file consisting of features from datasets collected at all geographical levels. This script follows the same guidelines for both time series data and non-time series data.

Baseline Forecaster

The baseline forecaster (BF) projects historical time-series data into the future (every year from current data to 2040). Multiple data augmentation pipelines and forecasting models are recorded in the experiments. A detailed description of the experiments conducted for BF is included in Section 4. The performance of the deployed model is stated below:

a. Within 8% of Root Mean Squared Error (RMSE) for the next year
b. Within 6% of Mean Absolute Error (MAE) across all years
c. Regularized, with no unrealistic exponential increase in the future

[1] Linear interpolation for missing data between years. Backward and forward filling for missing data at the ends in a time series.

[2] Smaller administrative boundaries will inherit data from their larger administrative boundaries. For example, counties will inherit data only available at state level and zip codes will inherit data from counties.

Scenario Forecaster

The Scenario Forecaster (SF) aims to learn the relationships of different features contained in the features set and return scenario-based forecasts from an explainable machine learning model. The performance of the deployed model is stated below:

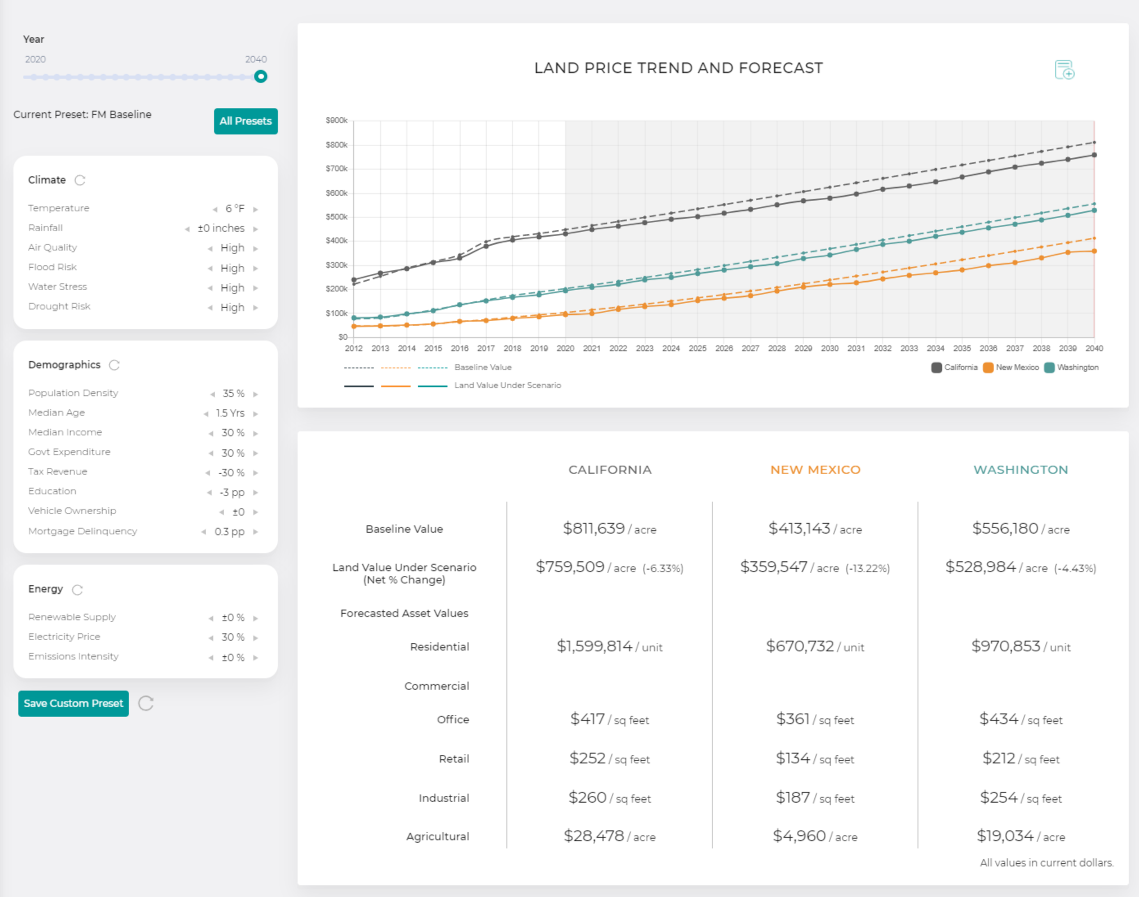
a. Within 6% of Root Mean Squared Error (RMSE)
b. Within 4% of Mean Absolute Error (MAE)
c. Sensitive to variations in input conditions.

Scenario Parser

Our models use ~50 features related to real estate markets, socio-economic conditions, environmental characteristics, and many others to estimate asset values.

The Scenario Parser (SP) enables scenario-based forecasting by modifying the data stream fed to the SF. For example, if a user increases coastal flood risk, the features that correspond to coastal flood risk will be adjusted accordingly by the SP and a new input condition will be used for downstream inference.

To create a scenario, users are allowed to adjust variables within each scenario preset as input for the scenario parser. Figure 2 shows a custom scenario example created by a user with different modifiers applied to the baseline scenario. Users can input any combination of updates to variables to feed into the scenario parser. The input of the user is sent to the scenario parser and the parser modifies the input data to the forecaster. The scenario forecaster then updates the price trends based on the modified input data.



Current Preset: FM Baseline

Climate

Temperature	6 °F
Rainfall	±0 inches
Air Quality	High
Flood Risk	High
Water Stress	High
Drought Risk	High

Demographics

Population Density	35 %
Median Age	1.5 Yrs
Median Income	30 %
Govt Expenditure	30 %
Tax Revenue	-30 %
Education	-3 pp
Vehicle Ownership	±0
Mortgage Delinquency	0.3 pp

Energy

Renewable Supply	±0 %
Electricity Price	30 %
Emissions Intensity	±0 %

	CALIFORNIA	NEW MEXICO	WASHINGTON
Baseline Value	$811,639 / acre	$413,143 / acre	$556,180 / acre
Land Value Under Scenario (Net % Change)	$759,509 / acre (-6.33%)	$359,547 / acre (-13.22%)	$528,984 / acre (-4.43%)
Forecasted Asset Values			
Residential	$1,599,814 / unit	$670,732 / unit	$970,853 / unit
Commercial			
Office	$417 / sq feet	$361 / sq feet	$434 / sq feet
Retail	$252 / sq feet	$134 / sq feet	$212 / sq feet
Industrial	$260 / sq feet	$187 / sq feet	$254 / sq feet
Agricultural	$28,478 / acre	$4,960 / acre	$19,034 / acre

All values in current dollars.

Figure 2

Scenario Modulator

The Scenario Modulator applies a precalculated price coefficient based on statistical models as an additional influence on the forecasted price generated by the Scenario Forecaster (SF). While the SF can learn the correlations between features and the target variable, features such as those that have not been priced in by the market have little correlation with the targets. However, the fundamentals that determine property price will change in the future. For example, climate factors such as environmental risks, temperature, and rainfall currently do not have a strong bearing on property prices, indicating that these factors have not adequately been 'priced' into property valuations. However, this will change in the future as more investors begin to price in the impact of climate. Since we do not have future real estate data that have climate risks factored in, this cannot be tackled with a supervised ML model.

To incorporate the future shifts in fundamentals, a scenario modulator is created using coefficients calculated through various statistical models, and domain knowledge from experts in academia and the industry. The coefficients of the scenario modulator are regularly updated based upon industry research and consultation with experts.

4. BASELINE FORECASTER DETAILS

The Baseline Forecaster forecasts yearly data until 2040 from historical trends. In total, 99 features from the upstream data processor are forecasted. As our dataset is grouped, artificial neural networks are able to encode each geographical territory as individual ‘neurons’ and learn the trends of each territory while keeping the national and regional averages in mind, resulting in a more geographically correlated forecast into the future. To determine the best approach to forecasting the future, we experimented with different model architectures and data pipelines. We found that the best performing pipeline uses a Dense Neural Network (DNN) with an optional Convolutional layer predicting 12 timesteps into the future. Further details on experimental results can be found in the sections below.

Current Architecture

Figure 3 is a high-level description of the neural network used for baseline forecasting. It first comprises an embedding layer that groups the time series of different geographical regions into sub-series that are specific to one geographical region only. This region might be at the level of State, County, CBSA, Zip Codes, or other geographical areas of concern.

The embedded data layer is connected to a data augmentation layer that up-samples the number of data points of the embedded inputs while maintaining the historical trend. This layer helps to regularize the prediction and improve downstream accuracy.

The upsampling layer is then connected to a 1-Dimensional Convolutional layer with the appropriate kernel size which is in turn connected to 2 or 3 layers of hidden fully connected layers after applying a pooling process.

Afterward, a downsampling layer reduces the data points generated by the upsampling layers into the original structure of the time series. The final output layer returns the forecasted results of the specified variable in the next year, and the entire neural network is built in an autoregressive loop where the predictions are fed into the network until the desired length of forecasted results is reached.

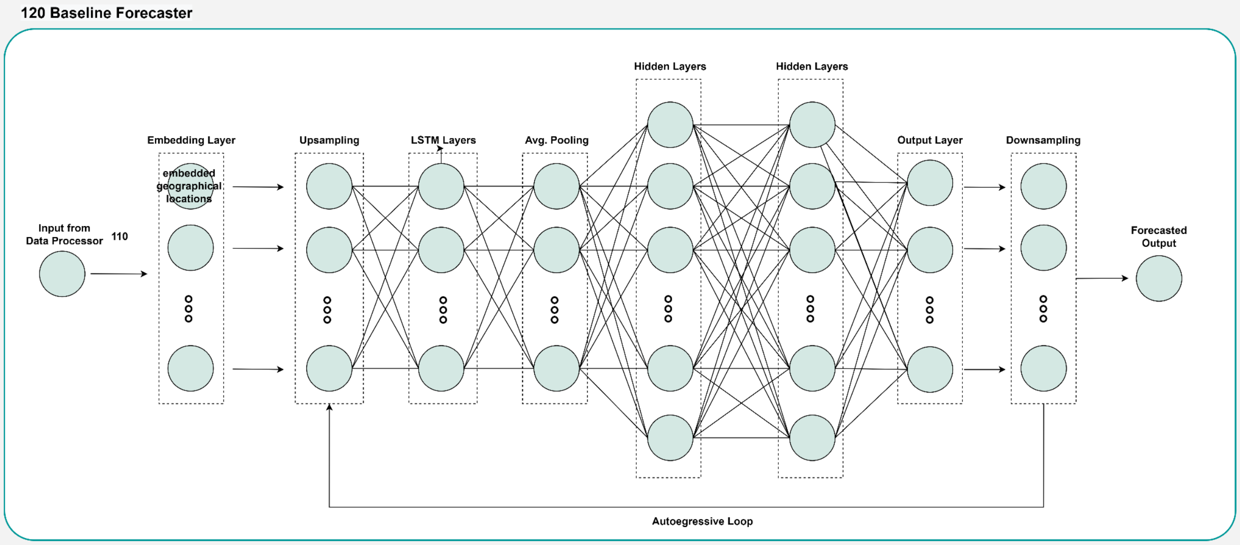


Figure 3: Baseline Forecaster Architecture Diagram

Experiment Setup

This section presents an overview of notable iterations in past experiments on the baseline forecaster which led to the final selected model described above.

Baseline	The baseline is simply a linear regression model in the previous year.
ARIMA	Autoregressive integrated moving average (ARIMA) is a classic statistical model widely used in forecasting. This serves as an alternative baseline for performance measurement.
Dense	Facebook's Prophet is a popular classical machine learning forecasting library used in many projects in practice. This serves as a performance reference to the off-the-shelf ML models.
CNN	A custom ANN with 2 dense layers with 128 nodes using ReLu activation function.
Random Forest (RF)	Two 1-dimensional convolution layers are added before the dense layers with variable kernel sizes in different iterations that involve 3 or 7 years of historic data for the model to learn historical patterns.
LSTM	Two bi-directional LSTM layers are added before the dense layers so that past year data can be 'memorized' and impact predictions as well.

Metrics

The performance of the models are measured through back-testing of past year data with k-fold validation. The total dataset of 1000 counties is split into 5-folds, with each fold having different geographical compositions.

We use Mean Absolute Error (MAE), Root Mean Squared Error(RMSE), and Mean Absolute Percentage Error(MAPE) to measure the performance of the models. The errors are calculated by rescaling the results back in dollar terms against the ground truth.

Index-to-price translation

One of the most challenging aspects of real estate prediction is the lack of historical data. Most of the data available from the 1970s to 2010s are in price indices, rather than transacted amounts. Datasets with transaction prices in dollar terms are mostly only available after the 2010s. This means that backtesting is only limited to data from recent years, and is unable to accurately predict property values across longer real estate cycles

To overcome this limitation, we have developed an index-to-price translation model that converts any real estate price index (e.g. FHFA Home Price Index, Commercial RE price index) into absolute dollar terms, within a 2.4% mean absolute percentage error across all counties and years.[3] This method expands the temporal coverage of the transacted price dataset and enables more accurate long-term predictions.

Results

The results of each model are recorded below. All the performances are derived from the same backtesting and pre- and post-processing techniques. Using backtesting on every past year, the Mean Absolute Percentage Error is ~ 6% and Root Mean Squared Percentage Error is ~ 8%.

[3] Please refer to Appendix experiments conducted to obtain the best approach.

Model	Key Layers and Hyperparameters	Train MAE($)	Train RMSE($)	Test MAE ($)	Test RMSE($)	Test MAPE	Test MSPE
Baseline	layers=(Dense(1),Dense(9)), activation='relu'	3080	6255	11994	26568	3.6	9.2
ARIMA	difference=2, p,q selected from grid search, using least AIC score.	4873	9166	11863	18725	4.1	6.6
Prophet	Epochs = 956, batch_size = 16, learning_rate=0.04	3854	6631	16725	29926	5.2	9.8
Dense	layers=(Dense(512),Dense(9)), activation='relu'	3093	6449	11883	25501	3.6	9.1
CNN3	layers=(Conv1D, Dense), Kernel_size=3	3019	6551	11869	25985	3.6	9.1
CNN7	ayers=(Conv1D, Dense), Kernel_size=7	2662	5188	10354	20602	3.4	6.9
LSTM	layers=(LSTM, Dense), lstm_units=32	2913	5788	8903	14638	3.1	5.1

Key Discoveries

The model performs better against other baseline algorithms such as linear regression and ARIMA based on MAE and RMSE. By upsampling yearly data to monthly data, we have 'regularized' the future prediction trend and avoided unrealistic exponential growth forecasts. In datasets with high irregularities, applying moving averages greatly stabilizes the resultant trend and regulates the forecast.

This baseline data provides a good projection of the current trend into the future and can be used as the baseline for training the downstream scenario forecaster.

Limitations

While the chosen method has a low margin of error, the baseline prediction is only a forecast on the expected average trendline in the next two decades. It does not take into account fundamental shifts in market sentiments, macro-economic policies, or other exogenous variables such as climate change.

Future Improvements

1. **Rural-Urban Divide.** Some datasets (e.g. multi-family homes and commercial property prices) only cover urban areas (i.e. CBSAs). More data relevant to counties outside of the urbanized areas are being collected to examine the different drivers of rural and urban asset prices. In the future, urban areas and rural areas can be trained individually, increasing our coverage to all counties.

2. **Rate of change forecasting.** The current forecasting is based on absolute values. It is possible that forecasting index values will return even better results as the index values indicate the rate of change and have less skewness than the absolute dataset (less tail heavy), leading to better training performance.

3. **Ablation Study.** The best-performing model described above uses multiple data pre-processing steps to achieve the results. We will offer an ablation study to showcase the effectiveness of our data processing pipeline in reducing backtesting errors.

5. SCENARIO FORECASTER DETAILS

After obtaining a baseline forecast into the future from the baseline forecaster, the data is then used to train a scenario forecaster that responds to the additional conditions set by the user on the front end as seen in Figure 2.

Many different machine learning models were experimented to find the best model that offers accurate prediction and explainability. Figure 4 presents the current ML pipeline of the Scenario Forecaster.

Current Architecture

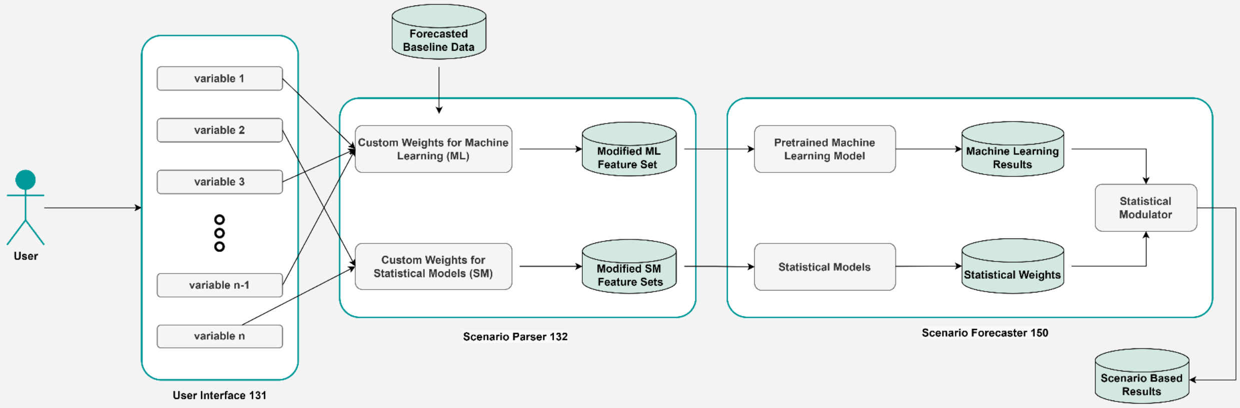


Figure 4: Scenario Forecaster Pipeline

The user interface enables the user to increase or decrease the likelihood of certain predefined scenarios, or to adjust the values of individual independent variables to be used for the generation of the scenario-based prediction as seen in Figure 2 in the previous section.

The scenario parser receives inputs from both the user interface and Baseline Data. Some or all of the forecasted baseline data may be displayed to the user via the user interface as a baseline scenario. The baseline scenario is a projection of the future values of the plurality of variables and can be adjusted by the scenario parser in accordance with the input from the user interface. Accordingly, the scenario parser outputs modified values of the plurality of variables.

The scenario parser processes the user input, reads the user input, and encodes the input into different weights to adjust the input dataset for prediction. It receives the baseline data and separates it by the features learned by the machine learning model from features outside of the model. Features that are learned by the machine learning model have different degrees of correlation or feature importance. Adjusting the value of the variables thus affects the machine learning results. Features that are not learned by the model are modeled with statistical models that are calculated based on research and heuristics.

The machine learning model comprises an ensemble machine learning model that is configured to estimate future values of the selected dependent variable based on forecast data for the independent variables. For example, the machine learning model may be obtained by training using the historical values of the independent variables and the forecast values of the dependent variables from the forecast data, and the historical values of the dependent variable and the forecast values of the dependent variable from the forecast data. In some embodiments, the machine learning model is obtained by retrieving stored parameters of the model from a database of previously trained models.

Experiment Setup

The goal of the scenario forecaster is to provide accurate predictions based on customized shifts in the input dataset, simulating a change in environment. The best model would be one that has a low error margin while maintaining the high explainability and sensitivity to high dimensional problems. The ML models that were chosen for experimentation are listed below. Compared to the Baseline Forcaster, the models do not use deep neural networks (DNN) as they generally have poor model explainability.

Linear	A linear regression model serves as a baseline and is easily explainable.
SVM	Support Vector Machines (SVM) are effective in high-dimensional spaces.
KNN	k-nearest neighbors (KNN) could cluster locations with similar traits.
Decision Trees (DT)	Simple tree-like regression model, a baseline for DT-based models.
Random Forest (RF)	An ensemble DT model with bagging. Easily tunable and generally returns more accurate predictions in higher dimensions problems than linear models.
XGBoost	An ensemble DT model with gradient boosting. Requires less hyperparameter tuning than RF and is more accurate with an unbalanced dataset.

LightGBM	In XGBoost, trees grow depth-wise while in LightGBM, trees grow leaf-wise which is potentially faster in training and deployment while maintaining similar performance.
Gradient Boosting(GB)	An ensemble DT model with Histogram-based Gradient Boosting. This estimator has native support for missing values (NaNs)

Spatial Encoding

As real estate is greatly affected by location, different experiments were conducted to find the best geographic encoding techniques. Currently, each county is encoded by a one-hot State code and ordinal encoded County code. This is to make sure counties in different states are separated into different feature spaces whereas counties in the same state can be clustered together in an ordinal series to capture their spatial similarity.

Metrics

Similar to the Baseline Forecaster, the performance of the models is measured through back-testing of past year data with k-fold validation. The total dataset of 1000 counties is split into 5-folds, with each fold having different geographical compositions.

We use Mean Absolute Error (MAE), Root Mean Squared Error(RMSE), and Mean Absolute Percentage Error(MAPE) to measure the performance of the models. The errors are calculated by rescaling the results back in dollar terms against the ground truth.

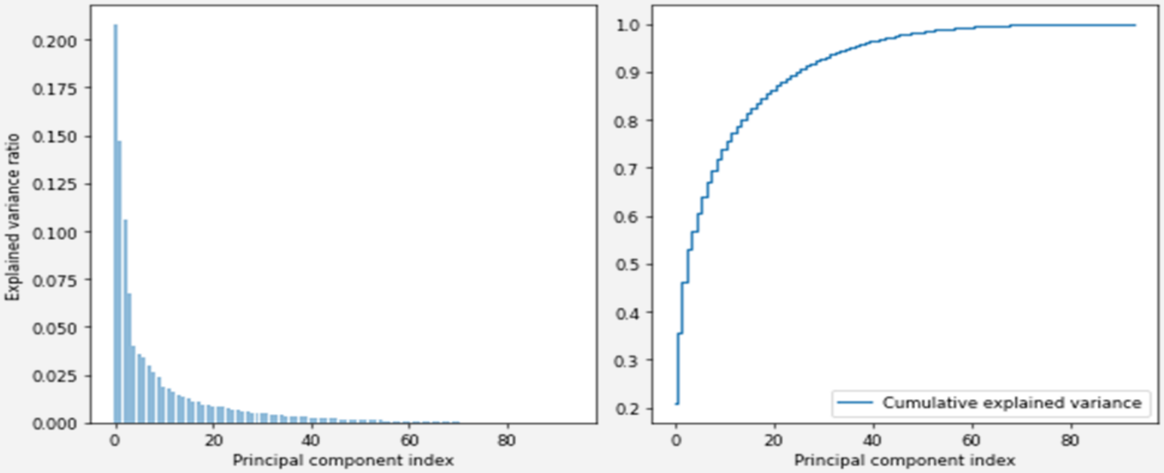
Results

The results of each model using 5-fold cross-validation are recorded below. All the performances are derived from the same backtesting and pre-and post-processing techniques with 72 features. From the results, we see that tree-based models such as Random Forest or Gradient Boosting offer superior performance compared to other regressors. Amongst the tree models, Random Forest (RF) performs the best in terms of test metrics with a MAPE of 7.1% and MSPE of 8.1. Gradient Boosting (GB) and LightGBM (LGBM) also offer good results with much lower inference speed, which is a crucial element in real-time usage.

Model	Train MAE ($)	Train RMSE ($)	Test MAE ($)	Test RMSE ($)	Test MAPE (%)	Test MSPE (%)	Inference Speed (ms)
Linear	53553.78	101995.55	55054.88	128219.45	23.6	36.1	4.8
SVM	210498.57	294286.80	206719.52	304510.39	**64.6**	**85.8**	**4.6**
KNN	27157.84	50776.29	25697.31	53743.86	11.0	15.1	6005
Bayesian Ridge	53432.82	101970.81	54984.22	128299.53	23.5	36.1	8
Decision Trees	31308.27	52973.49	29540.65	52918.23	11.2	14.9	5.6
Random Forest	18915.02	34533.74	17707.94	28742.48	**7.1**	**8.1**	114.8
XGBoost	54191.59	80368.51	54887.19	79700.87	23.2	22.5	28.3
Gradient Boosting	19456.63	38381.76	19607.51	44471.26	7.9	12.5	33.8
Light GBM	19161.71	38774.48	18828.34	50219.63	7.7	14.1	23.4
RF + GB	20245.48	36587.65	19787.691	33981.39	7.8	9.5	**283.2**

Feature Optimization

Next, experiments are conducted with **RF, GB and LGBM.** The results of these models can be optimized further through feature optimization and hyperparameter tuning. While Climate Alpha has a rich pool of datasets, further feature engineering is conducted on the initial feature space to trim away unimportant or noisy features that could hinder the performance of the model or muddle the explainability of the models.



Explained Variance ratio and cumulative explained variance

Using principal component analysis, we found that approximately 50 features out of the entire set of 99 features contribute to 95% of the explained variance. This means that many features offer similar correlations and can be removed without impacting actual model performance. Through exploratory data analysis (EDA), 27 features were dropped from the initial set of 99 features. Furthermore, 7 climate variables were found to have little correlation with the target variables. Since the impact of climate variables is expected to change in the future, their historical impact does not contribute to the performance of the model. This reduced the feature count to 65. Lastly, combining analysis results from explained variance ratio from PCA and feature importance values from the selected models, 15 features were dropped again, leaving the final feature size at 50. The following table shows the performance of 5-fold CV results of the selected models trained on decreasing feature sizes. Notice that performance did not vary drastically, therefore confirming that the optimized feature set can effectively explain the correlation with the target variable just as well as the initial dataset. All metrics are calculated on the test set.

Dataset (Features)	**Model**	**Key features**	**MAE ($)**	**RMSE ($)**	**MSPE (%)**	**MAPE (%)**	**Inference Time (ms)**
Initial Features (99)	RF	45	18925	34220	9.6	7.5	80.7
	GB	40	19878	32561	9.1	7.9	34.1
	LGBM	52	19617	32951	9.2	7.8	20.2
Reduction after EDA (72)	RF	35	19030	39333	11.0	7.6	77.3
	GB	50	20732	35898	10.0	8.1	30.0
	LGBM	52	20639	32690	9.13	8.3	19.2

Reduction of Climate Variables (65)	RF	31	19585	43221	12.1	7.7	53.1
	GB	45	21791	37473	10.5	8.8	31.4
	LGBM	49	21091	34600	9.6	8.4	19.2
Final Features (50)	RF	30	19525	38931	10.9	7.7	80.0
	GB	40	22444	36366	10.2	8.9	30.0
	LGBM	45	22010	34632	9.6	8.9	19.0

Hyperparameter Tuning

Using the final feature set, the hyperparameters of the three candidate models are tuned. The LGBM model with the lowest error and relatively small model size is chosen for deployment. Their final performance is shown in the table below:[4]

Model	Test MAE ($)	Test RMSE($)	Test MAPE(%)	Test MSPE(%)	Inference Speed (ms)	Model Size (mb)
LGBM	9447.48	16700.15	3.66	4.66	80	6.2
GB	12405	21903	4.82	6.12	50	3.2
RF	19525	38931	7.7	10.9	80	349.4

Limitations

1. **Shifting Baselines.** While we have achieved good results after iterating through different experiments, is it important to note that there is no perfect foresight. Backtesting, however rigorous, does not account for black swan events or sudden shifts in government policies. As such, the model will be continually re-trained as soon as new records of its features are updated.

2. **Relative rather than absolute**. It is also important to note that the results given by the SF are indicative rather than absolute. The deviation from the baseline trend helps investors to be on the correct side of the macro trend, rather than pinpointing the price of assets.

[4] Please see the appendix for the chosen hyperparameters.

Future Improvements

1. **Encoding by climate classifications**. To increase geographic sensitivity further, we will encode each county with their respective Köppen Climate Categories to account for climate differences between different geological features even amongst counties that are close to each other. For example, some counties could be separated by an important geological feature that creates a different microclimate in the two counties. Encoding Köppen Climate classifications would prove the model of this layer of information.

2. **Offering other models.** While the current model has the lowest error, other models can also be offered for users to explore different types of correlations to make more informed decisions. For example, a linear model might perform less well in backtesting but offers a more comprehensible correlation coefficient than the feature importance used in tree models.

6. FEATURE WEIGHT EXPLORER

Aimed at enhancing the explainability of our results, the Feature Weight Explorer was created to explain the effect of the chosen scenario on land value, to be used in tandem with the scenario forecaster. The feature explorer converts the feature importance learned by the model into percentage points, giving users a more intuitive understanding of the impact of each variable. Figure 5 shows the impact on property price for each 1% increase in feature value in California. The figure also shows the projected impact in the future that is not yet priced in. This future projection is deduced from our statistical models which will be explained in the following chapter.

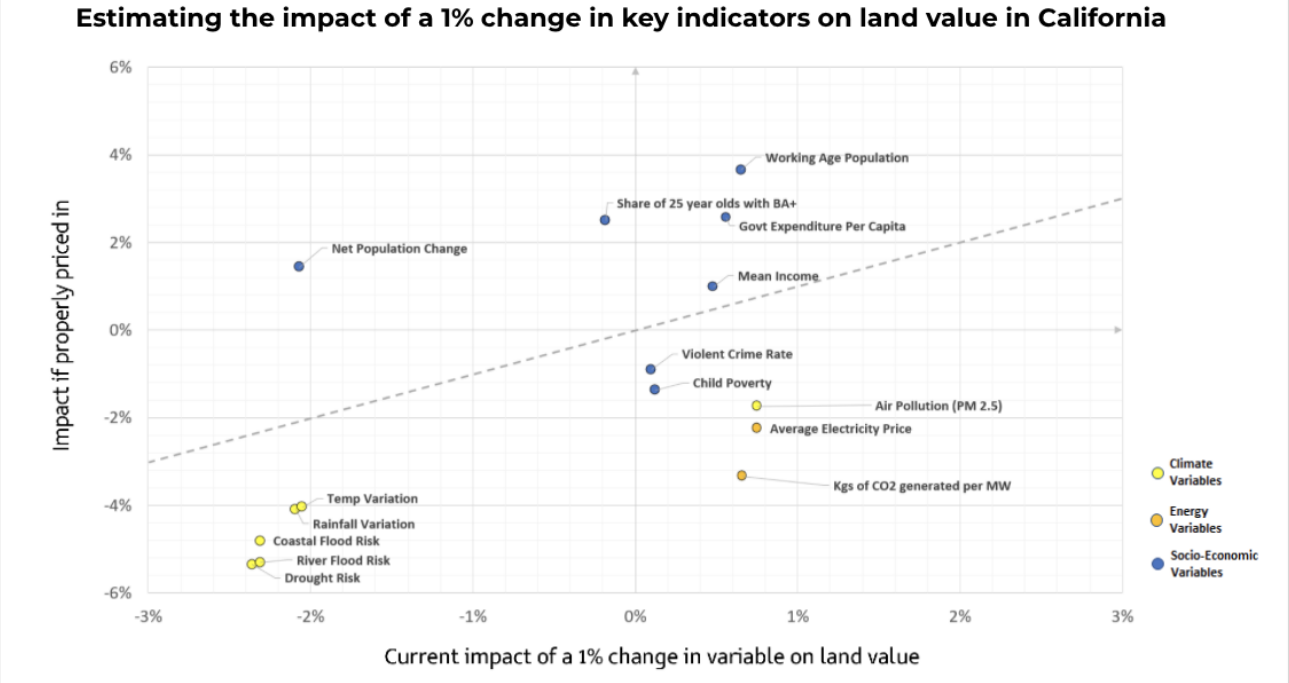
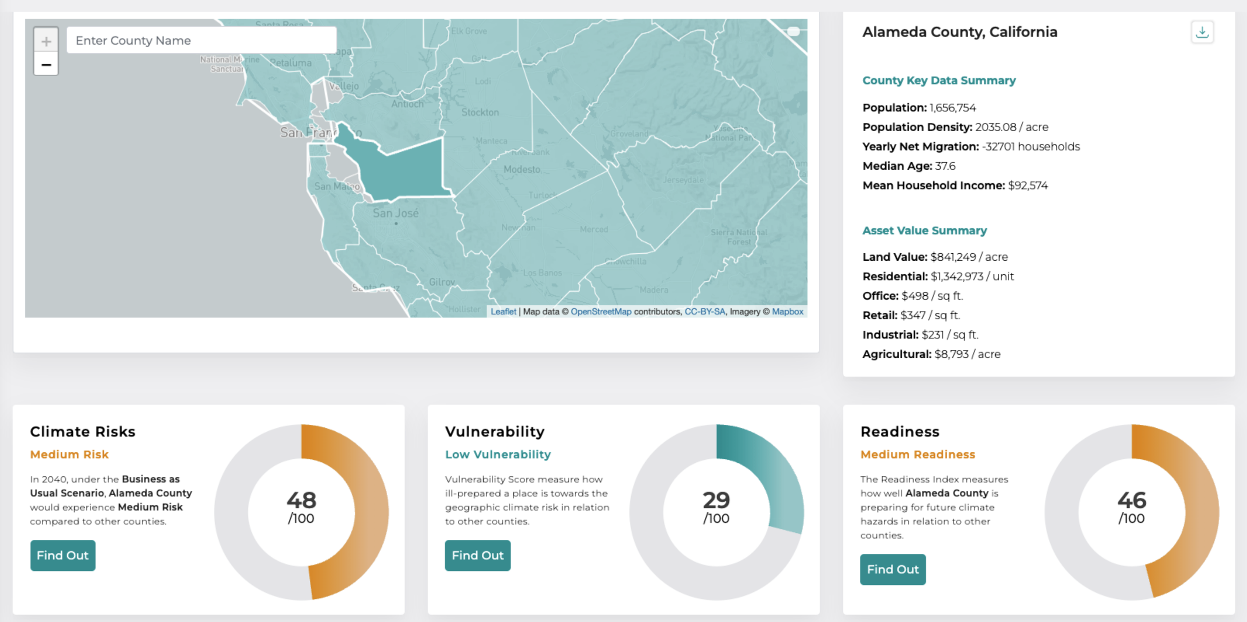


Figure 5: Feature Weight Explorer

This graph shows the impact of key variables on land price in California, with the current impact of a 1% change in a variable on land value per acre shown on the x-axis, and its likely future impact was shown on the y-axis. Notable likely corrections include emissions intensity (defined as kilograms of CO2 generated per MW) which is currently positively correlated with land value. We expect that as climate risks become more pronounced, land prices in geographies that rely on fossil fuels will see an increasingly strong negative correlation with emissions intensity. For California, by 2040, we expect a 1% increase in emissions intensity to lower land prices by ~3%.

7. RESILIENCE INDEX ™



Alameda County, California

County Key Data Summary

Population: 1,656,754
Population Density: 2035.08 / acre
Yearly Net Migration: -32701 households
Median Age: 37.6
Mean Household Income: $92,574

Asset Value Summary

Land Value: $841,249 / acre
Residential: $1,342,973 / unit
Office: $498 / sq ft.
Retail: $347 / sq ft.
Industrial: $231 / sq ft.
Agricultural: $8,793 / acre

Climate Risks

Medium Risk

In 2040, under the **Business as Usual Scenario**, **Alameda County** would experience **Medium Risk** compared to other counties.

48 /100

Vulnerability

Low Vulnerability

Vulnerability Score measure how ill-prepared a place is towards the geographic climate risk in relation to other counties.

29 /100

Readiness

Medium Readiness

The Readiness Index measures how well **Alameda County** is preparing for future climate hazards in relation to other counties.

46 /100

CA Resilience Index Dashboard for Alameda County, California

Background

While the machine learning models in the previous sections offer insight into the features that drove the property market historically, the future market will likely be based on an even more complex set of fundamentals than today. The most perceivable shift in fundamentals is the rising importance of climate and sustainability-related factors such as climate change-induced hazards (physical risks) or ESG related costs (transitional risks). Although these features have had a low historical correlation with property values, rising flood, fire, coastal risk and other insurance premiums indicate that property markets are beginning to price in these factors.

The Climate Alpha Resilience Index ™ was crafted to provide:

(i) An intuitive index for qualitative analysis of climate change-induced impact across geographies, and

(i) A statistical model to account for future shifts in market fundamentals cause by climate change, by assigning weighted coefficients to the machine learning output based on different climate change scenarios.

Overall Scoring Methodology

The main index is aggregated from three sub-indexes, modeling Climate Risks, Climate Vulnerability, and Climate Readiness separately for ease of qualitative understanding. The index considers future climate projections, exposure to extreme events (e.g. population density and percentage area exposed to flooding), the sensitivity of the county population towards climate risks (e.g. percentage of the elderly population), and the capacity to cope with risk events (e.g. hospital beds per 1000 population).

The coefficients of the climate risk models are computed based on the output of our in-house climate risk models (Climate Alpha Risk Models) which are composite scores built upon extensive historical and statistical analyses of various GIS and climate data. In addition to risk models, we also compute a readiness coefficient for each county to gain insight into how prepared the counties are against the projected risks. This holistic scoring approach factors both climate risks and readiness in one analytic platform.

The Resilience Index is calculated from all continental U.S. zip codes and counties using a min-max approach, standardizing the performance of each location under each indicator on a scale of 0 to 100. Five categorical scores (i.e. Very low, low, medium, high, very high) are also assigned to each indicator based on a quintile split. The overall score of each category (Risk, Vulnerability, Readiness) is an average calculated from the respective indicator sub-scores. Each indicator score is weighted evenly in the aggregation. As some indicators are made up of more than one data source, in which case the multiple data sources are also evenly weighted to avoid aggregation biases.

Risk Index Methodology

The risk index considers future climate projections, exposure to extreme events (e.g. population density and percentage area exposed to flooding). Overall, climate risk in our model can be expressed as such:

Climate Risk =
Hazard Exposure x ***Hazard Likelihood*** x ***Deviation from Historical Thresholds***

Each climate change-related risk is aggregated from a list of climate projections under different carbon emission scenarios. These projected features are then engineered into scores for different climate change-related risk categories. There are currently five risk categories and each of the categories are listed below with their constituent datasets.

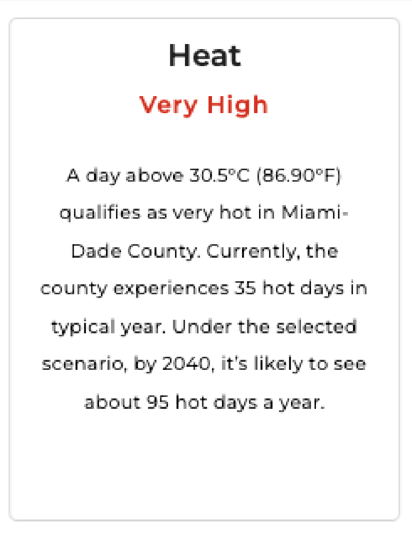
Heat Risk	Storm Risk	Drought Risk	Fire Risk	Flood Risk
Likelihood of heat event[5]	Likelihood of storm event[6]	Likelihood of drought event[7]	Historical fire events	FEMA Flood Score
Mean annual temperature	Storm intensity	Freshwater availability	Vegetation cover percentage	Surface runoff volume
Diurnal temperature range	Wind	Freshwater demand	Wind speed and humidity	

[5] Days with maximum temperature above 95th percentile based on history from 1960 to 2020.
[6] Number of continuous 2-day storms with daily precipitation above the 95th percentile.
[7] Number of continuous dry days below the 5th percentile.

Each location also has different risk thresholds calculated from historical data. For example, a day with a temperature above 86.9℉ is considered extremely hot for Miami, FL, while the threshold is reduced to 76.1℉ for Seattle, WA.

Taking the Resilience Index ™ flipcard to the right as an example, any day with a maximum temperature above 86.9℉ will cross the 95th percentile threshold based on historical climate data from 1960 to 2020. In 2040, the number of extremely hot days will increase from today's 35 to 95, which significantly increases the likelihood of heatwaves in the region.

An exponentially rising penalty is attributed to locations that cross their historical thresholds with greater frequency in order to account for potential tipping points or cascading effects.



Heat

Very High

A day above 30.5°C (86.90°F) qualifies as very hot in Miami-Dade County. Currently, the county experiences 35 hot days in typical year. Under the selected scenario, by 2040, it's likely to see about 95 hot days a year.

Climate Projection Scenarios

It is common practice to create climate projections under different emission pathways. Currently, there are different projections based on different assumptions and emission models such as Representative Carbon Pathways (RCP) or Shared Socioeconomic Pathways (SSPs). Climate Alpha gathers projection data from a variety of these models and presents 3 intuitive classifications of the future projections. These three scenarios are **Optimistic, Business as Usual, and Pessimistic**. The following table describes each scenario.

Optimistic (SSP1 RCP3.4)	The Optimistic scenario (SSP1 RCP3.4) represents a world with stable economic development and carbon emissions peaking and declining before 2040, with emissions constrained to stabilize lower than ~650 ppm CO2 and temperatures between 2.0 to 2.4°C by 2100.
Business as Usual (SSP3 RCP4.5)	The Business as Usual (BAU) scenario (SSP3 RCP4.5) represents a world with stable economic development and carbon emissions peaking and declining by 2045, with emissions constrained to stabilize at ~650 ppm CO2 and temperatures from 2.6 between 3.2°C by 2100.
Pessimistic (SSP5 RCP8.5)	The Pessimistic scenario (SSP5 RCP8.5) represents a fragmented world with uneven economic development, higher population growth, lower GDP growth, a lower rate of urbanization and steadily rising global carbon emissions, with CO2 concentrations reaching ~1370 ppm by 2100 and global mean temperatures increasing between 2.6 to 4.8°C relative to 1986–2005 levels.

Vulnerability and Readiness Index Methodology

The vulnerability and the readiness indexes measure how exposed or how prepared a location is to the physical risks, respectively. This information helps users assess the actual degree of damage of physical risks on their property, factoring in the climate resilience of their geography.

Vulnerability Index

The vulnerability index measures how exposed a place is to physical risks induced by climate change. It comprises 6 different sub-scores, covering the demographic, economic, and infrastructure vulnerabilities of a place.

Scores	Description	Score Composition
Population Density	Densely populated areas means that climate hazards would affect more population.	1. Population per acre
Coastal Population	Population near coastal areas has increased exposure to coastal flooding.	1. % of the population near coast
Age Structure	Elderly are more sensitive to sudden strikes of climate hazards such as heatwaves, heavy storms, and forest fires.	1. % population above 65 years old
Urban Porosity	Urban areas with low porosity are more susceptible to Urban Heat Island effect and suffer higher surface runoff during storms.	1. % built-up area 2. % road cover
Infrastructure	Areas with a higher number of older buildings and ill-maintained infrastructures are more sensitive to climate hazards.	1. % buildings built before 2000 2. % bridges and roads in poor condition
Poverty and Inequality	Impoverished families are more sensitive to the impact of climate hazards.	1. % population in poverty 2. % income spent on rent

Readiness Index

The readiness index measures how prepared a place is to physical risks induced by climate change. It comprises 6 different sub-scores, covering the demographic, economic, infrastructure, and governance readiness of a place.

Scores	Description	Score Composition
Credit Score	Credit score indicates the economic stability of a county.	1. Median FICO score 2. Debt to income ratio
Healthcare	Better healthcare resources increase adaptive capacity.	1. Hospital bed per 1,000 population 2. % population with health insurance
Crime and Safety	Lower crime rate hints at more income equality and stability in a county.	1. Violent crimes per 100,000 residents
Education	A well-educated population is more informed to make more climate-friendly decisions.	1. % population with high school education and above
Public Spending	Higher spending on infrastructure and public spaces increases climate readiness.	1. Per capita infrastructure spending 2. Per capita park spending
Clean Energy	Higher Clean Energy score signals the progress of transition into a more sustainable future.	1. Energy from renewable sources 2. Per capita clean energy investment

8. CLIMATE PRICE ™

Climate Alpha's Climate Price ™ is defined as the deviation from the baseline forecasted market value of a location, portfolio, or property when modeled under different climate change scenarios.

The Climate Price ™ is calculated as a percentage weight applied to the baseline forecast. This coefficient is derived from the Resilience Index scores (risk, vulnerability, and readiness) of each location, thus factoring in both exposure to risks as well as adaptation capacity. The risk score carries a 50% weight in the coefficient while the vulnerability and readiness scores are weighted at 25% each. (Both socio-economic variables and climate factors exhibit extreme or tail-heavy distribution by which negatively impacted locations suffer greater erosion of value versus the gains experienced by positively scored locations.)

Statistically, the coefficient is calculated from the Z-score of each location. The higher the deviation from the national mean performance of each cluster (risk, vulnerability, readiness), the more drastic the impact of climate risk is on asset values. The maximum and minimum impacts are anchored to the estimated range of economic impact caused by climate change from peer-reviewed journal papers[8]. Locations with similar climate risk profiles may have different valuation outcomes due to their varying readiness scores. For example, locations with higher readiness scores than surrounding areas may gain in value compared to their neighbors.

The Climate Price ™ disambiguates correlation and causation factors. The first-order impact of climate variables on asset values can be misleading (such as damage from a tropical storm). We focus on second-order impacts of climate variables such as insurance premiums and population movement that have a verifiable impact on asset values.

[8] Papers reviewed include:

S. Hsiang, R. Kopp, A. Jina, J. Rising, M. Delgado, S. Mohan, D.J. Rasmussen, R. Muir-Wood, P. Wilson, M. Oppenheimer, K. Larsen and T. House, Estimating economic damage from climate change in the United States. *Science*, Vol. 356 Issue 6345, pp 1362-1369 (2017).
T. Carleton & S. Hsiang, (2016). Social and economic impacts of climate. Science.

T. Carleton, S. Hsiang, Social and economic impacts of climate, *Science*, Vol 353 Issue 6304 (2016).

S. Hsiang, R. Kopp, A. Jina, M. Delgado, J. Rising, S.. Mohan, R. Muir-Wood, D. J. Rasmussen, M. Mastrandrea, P. Wilson, K. Larsen and T. House, American Climate Prospectus: Economic Risks in the United States (2014).

APPENDIX

Index-to-price Translation

Accurate translation of price indices to absolute dollar price is crucial for the baseline forecaster and helps users understand asset values better. Multiple methods were experimented with, and their results are recorded below.

Technique	MAE($)	RMSE($)	MAPE(%)	MSPE(%)
$Abs_i / Abs_{base} * 100$	9301.47	16258.32	3.8	6.8
hpi_{base} + Yearly Δ = hpi_{base+1}	7661.38	19983.71	2.6	8.2
$\Delta hpi_{2020-i}/\Delta Abs_{2020-i}$	14131.79	201336.84	3.5	84.2
$\mathbf{Abs_{2020}:Abs_i/hpi_{2020}:hpi_i}$	**5738.62**	**10694.06**	**2.4**	**4.5**
Linear Regression with 5 data points	5989.96	14263.21	2.4	6.0
Linear Regression with 6 data points	4617.41	11024.17	1.8	4.6
Linear Regression with 7 data points	3889.06	8874.97	1.5	3.7
Linear Regression with 8 data points	**3653.10**	**8189.77**	**1.5**	**3.4**

Table A1: Index-to-price translation metrics with different methods

From the table, two methods provide reliable translation. The first uses the ratio between the increase of HPI and the increase in the absolute value of year against the year 2020. This method can be used for places with very little data as only one year of absolute data needs to be collected for accurate translation.

The second method is Linear Regression. As the number of available data points increases, the error lowers significantly. Therefore, for places that have more than 6 years of absolute value data, Linear Regression will be used.