

# Visual Analytics Tools for the Global Change Assessment Model

Ross Maciejewski

Arizona State University

## GCAM Simulation

- After running thousands or even hundreds of simulations through GCAM this process generates an ensemble of datasets that are impossible to go through individually.
- Each dataset has been generated using varying parameters which may have minimal impact on the outputs while others may have a more drastic impact.

# Data Preprocessing

- In order to tackle the issue of processing this data it was necessary to define a file format that was flexible and would enable the extraction of key features from the datasets.
- JSON was the chosen file format with Scenario specific:
  - Input identification
  - Output query selection
  - Global feature extraction
  - Period of simulation

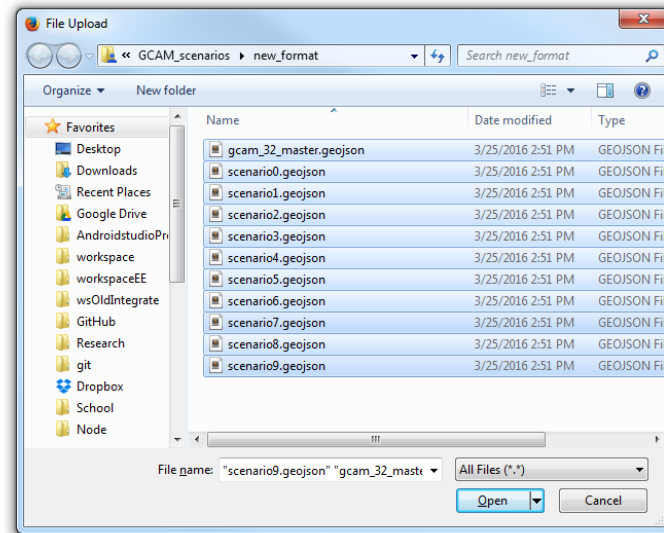
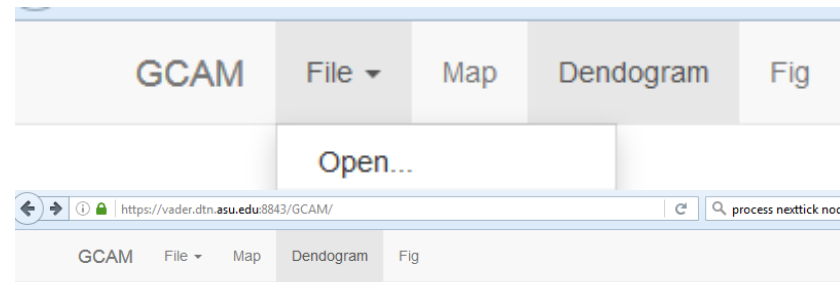
```
{
  "GCAM_ID": 13.0,
  "REGION_NAME": "EU-15",
  "GeoJSON_ID": 0,
  "Area": 10220.8,
  "queries": {
    "RCP_forcing": {
      "forcing-RCP" : ["1.5151", "...", "7.24347"],
    },
    "primary_energy": {
      "Oil": ["22.08653", "...", "20.6299"],
      "Nuclear": ["4.497564", "...", "4.558674"],
      "...": ["0", "...", "0"]
    }
  },
  "inputs" : {
    "Energy": "0",
    "CCS": "0",
    "SoEcon": "1",
    "AGLU": "0",
    "Ind/Tran/Build": "1",
    "Fossil": "1"
  },
  "years": ["1990", "...", "2100"]
}
```

## System Overview

- Our system is able to process these JSON files that contain:
  - Variable
    - Input parameters
      - The description of inputs that drove the simulation
    - Output parameters
      - The queries selected to be run in the simulation at the regional level
    - Global Parameters
      - The queries selected to be run in the simulation at the global level
- Providing analysis across a range of varying scenario

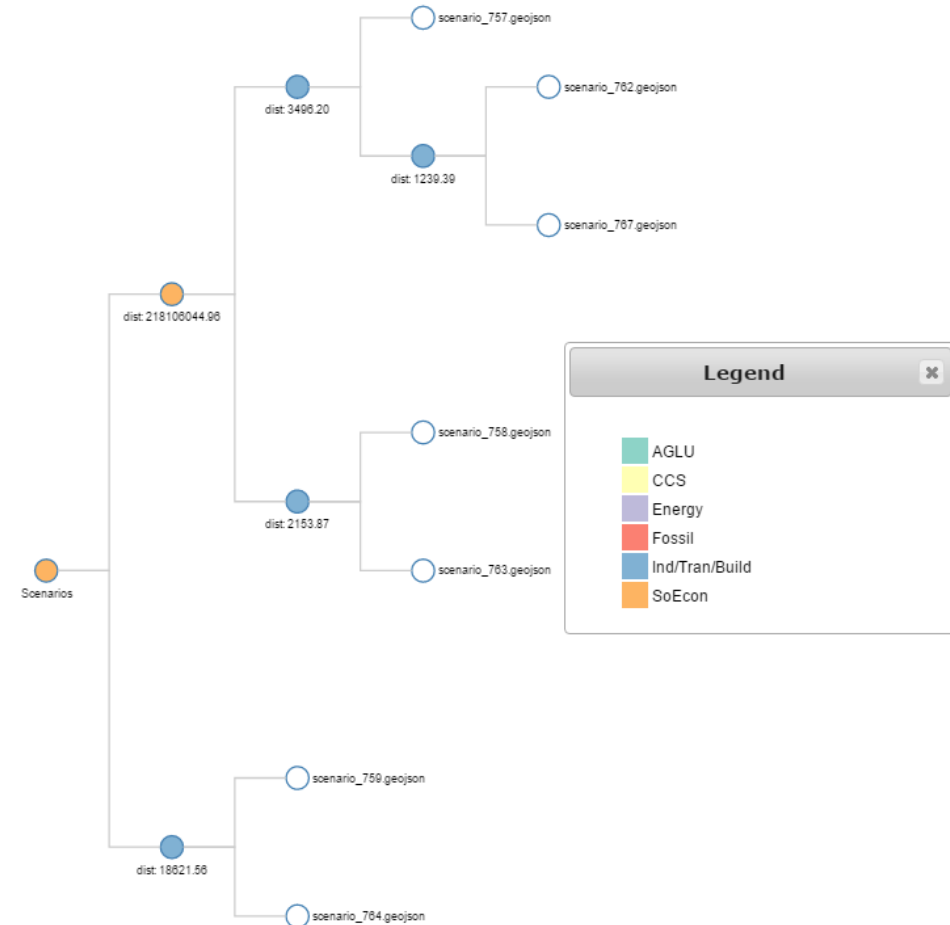
## System Implementation

- Desktop Application (Windows, Mac, & Linux)
  - Electron Framework
    - User Interface
      - Web Technologies
        - » HTML
        - » CSS
        - » JavaScript
    - Data Processing
      - NodeJS
      - Python



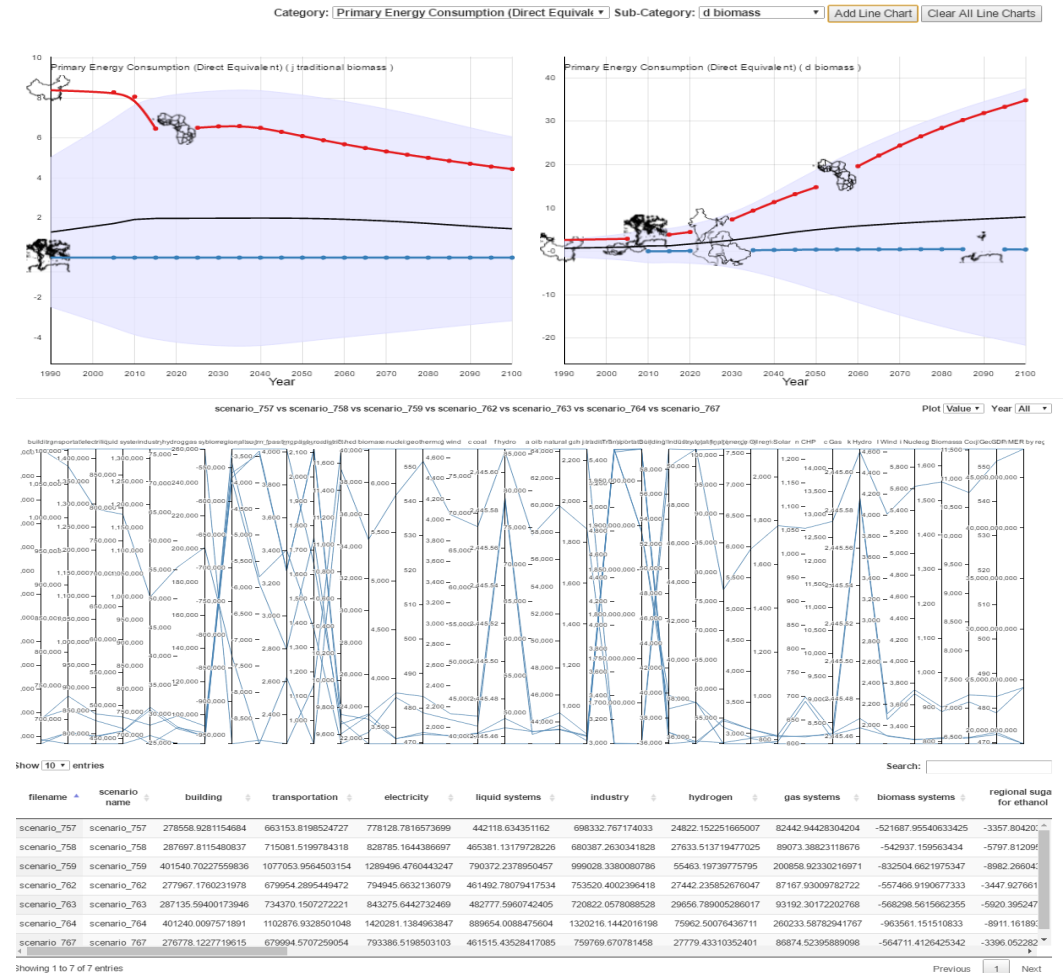
# Visual Analytics Techniques

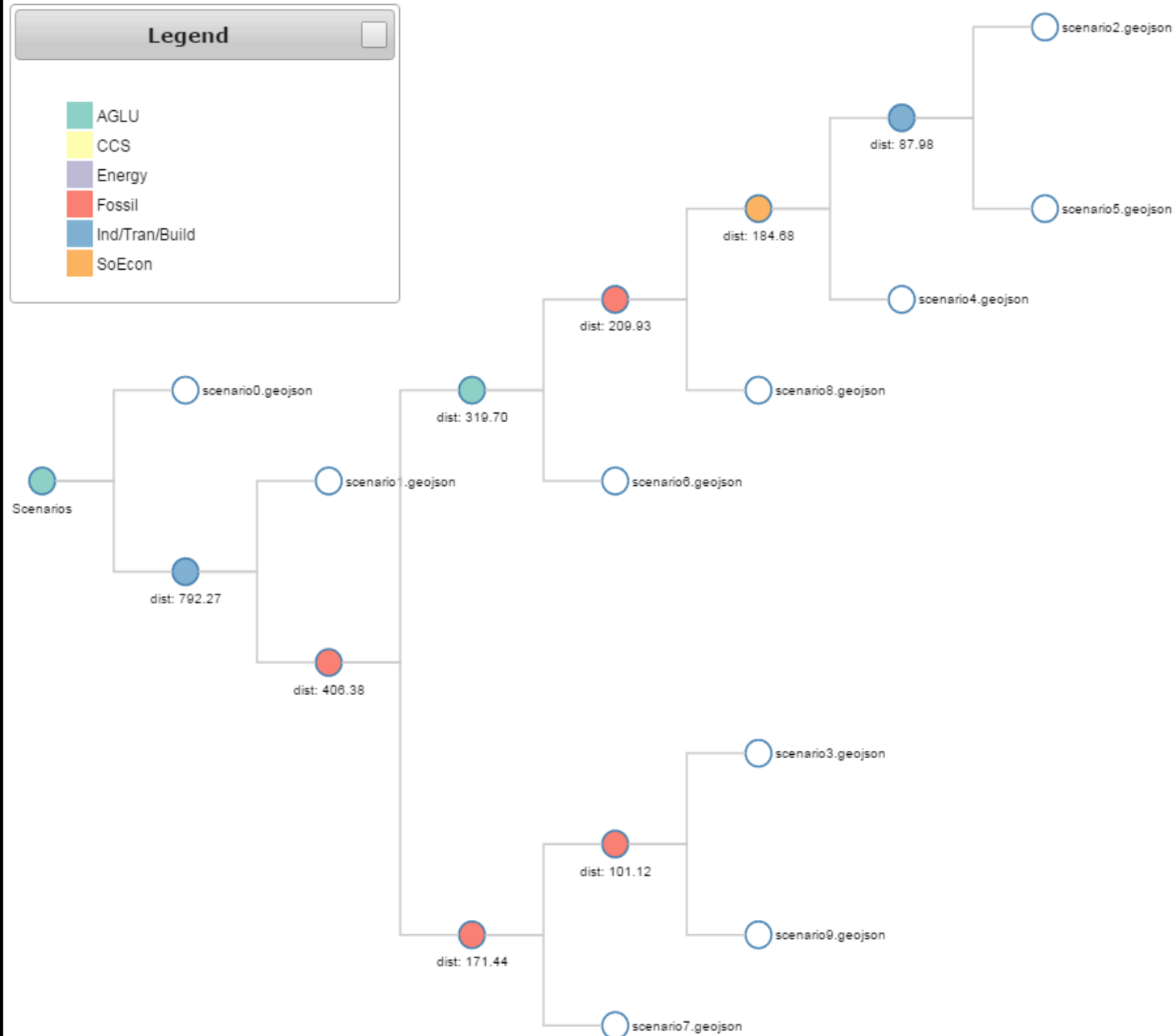
- Clustering
  - Hierarchical
  - K-Means
- Principal Component Analysis
- Correlation Identification
- Anomaly Detection



## Visual Analytics Views

- Dendrogram View
- Parallel Coordinates Plot
  - Scenario View
  - Difference View
  - Scenario Comparison View
  - Region Comparison View
- Min/Max View
- Cluster View



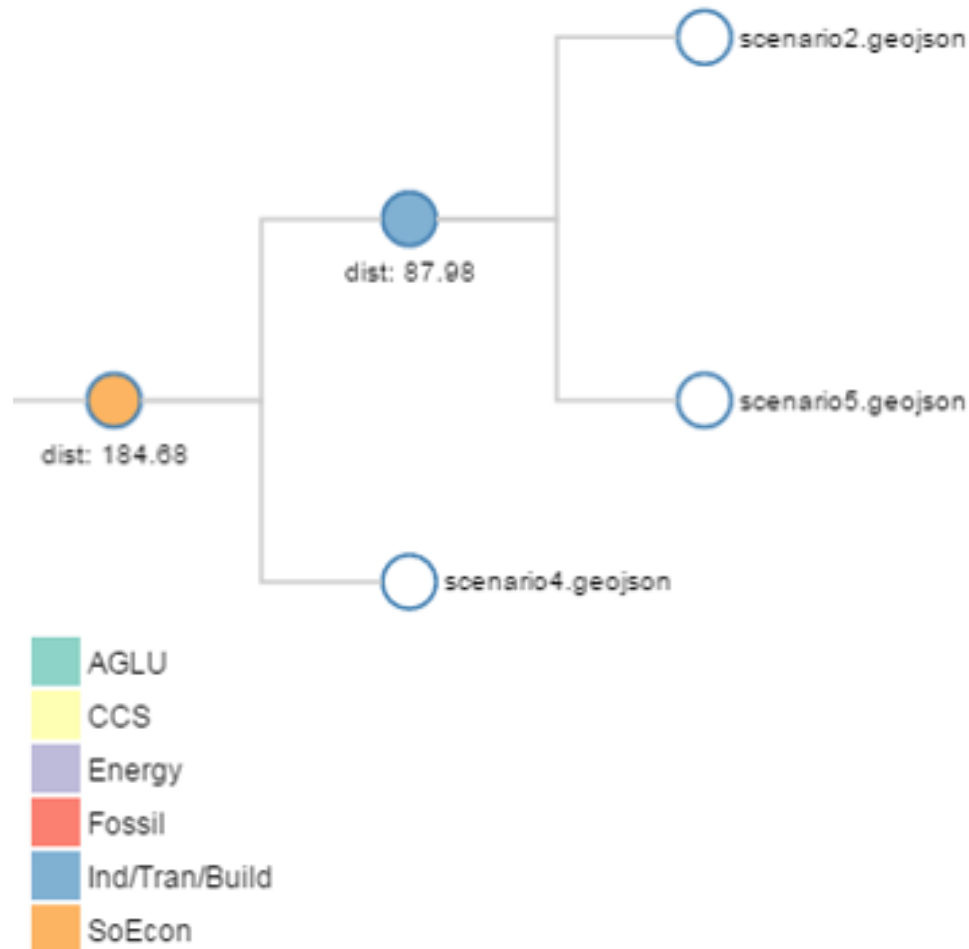


## Dendrogram View

- The dendrogram view shows the hierarchy of clusters formed when running when running all of the provided scenarios through hierarchical clustering.
- The view is draggable and zoomable allowing the user to easily explore the hierarchy.
- Nodes are clickable providing filter mechanisms for the Parallel Coordinates View and the Line Chart View.

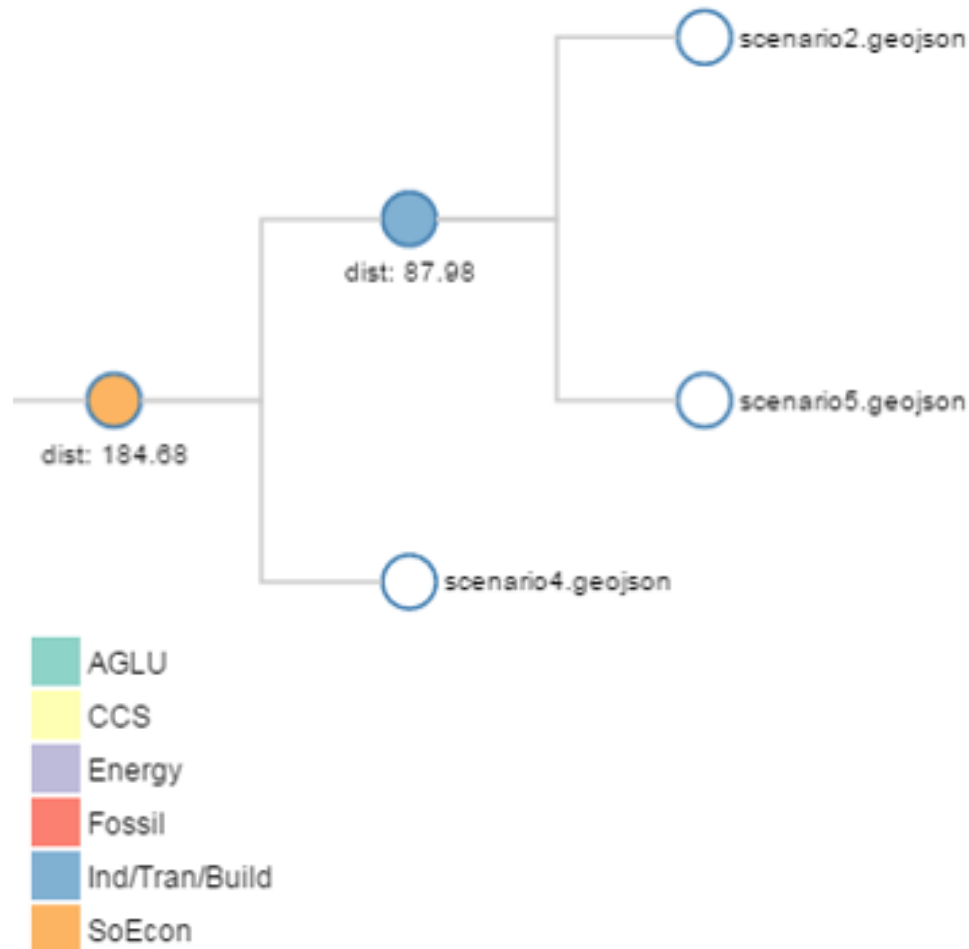


## Dendrogram View Cont.



- In the dendrogram, leaf nodes represent the scenarios and parent nodes represent the grouping procedure of clustering process.
- Distance between the nodes or clusters is labeled as 'dist'.
- To help users gain a better understanding of the clustering results, we label the parent nodes with the input parameter that has the largest difference at that step in the grouping process.

# Node Labeling Example



Scenario 2 inputs: Energy: 0, CCS: 0, SoEcon: 1, AGLU: 0, **Ind/Tran/Build: 0**, Fossil: 1

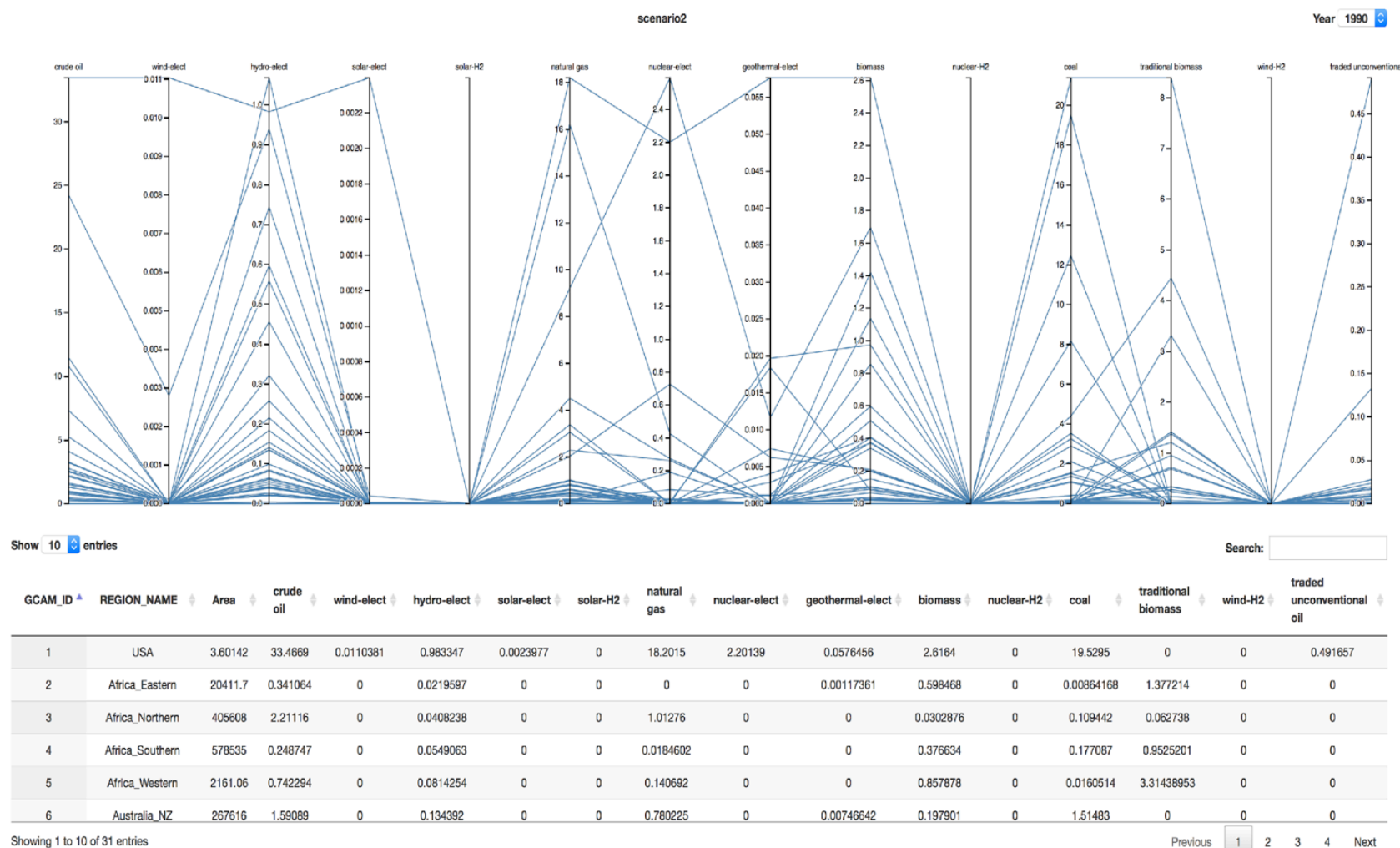
Scenario 5 inputs: Energy: 0, CCS: 0, SoEcon: 1, AGLU: 0, **Ind/Tran/Build: 1**, Fossil: 1

By applying the difference equation below we find the greatest difference between Scenarios 2 and 5 to be Ind/Tran/Build so the parent is colored Blue.

$$Diff_P = \sum_{R \in P} Absolute\ Value(Cluster1_{P,R} - Cluster2_{P,R})$$

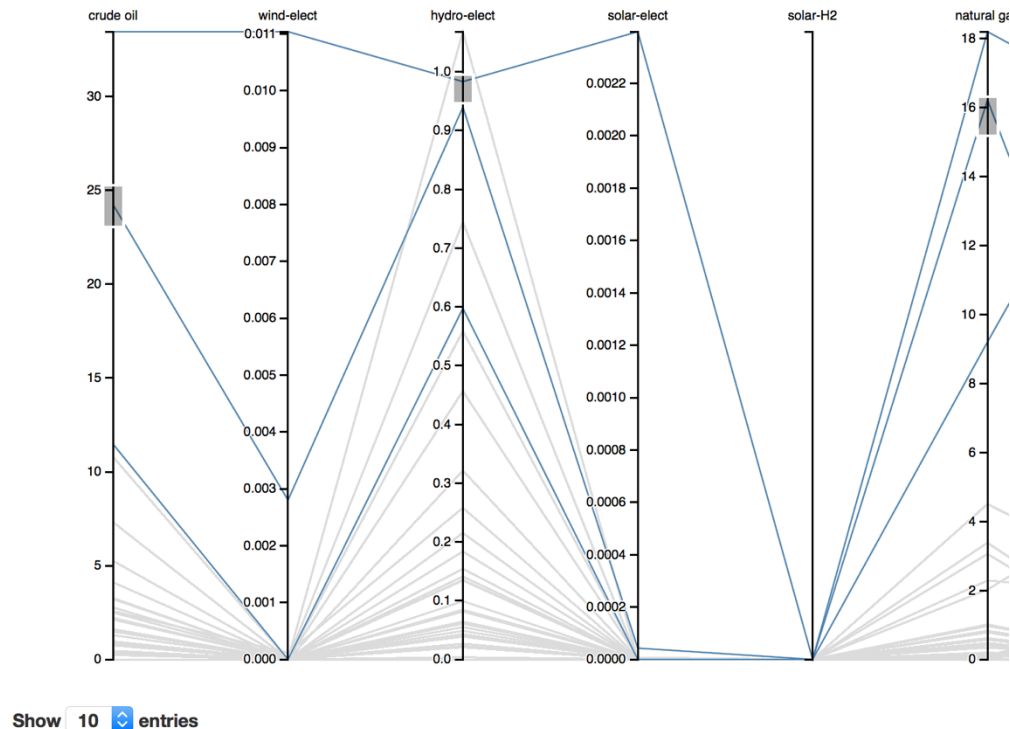
Where R is the Rth record of parameter P.

# Parallel Coordinates View



- Displays the output values for each region or scenario based on the selected queries.
- The plot can display the actual values and the slope of the values.
- The user can select to examine the data across all years as a summation or average, or they can select the value from a specific year in the simulation.

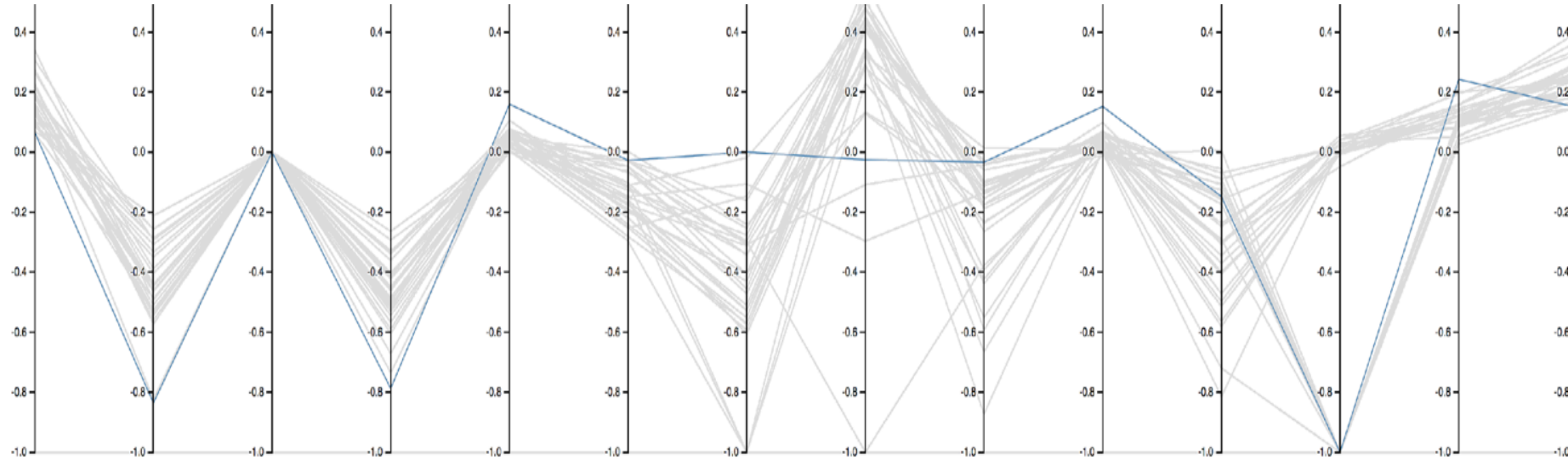
# Filtering options: Brushing



GCAM_ID	REGION_NAME	Area	crude oil	wind-elect	hydro-elect	solar-elect	so
1	USA	3.60142	33.4669	0.0110381	0.983347	0.0023977	
13	EU-15	10220.8	24.2373	0.00280083	0.939621	4.32005E-05	
23	Russia	16990100	11.4724	0	0.597301	0	

- Each y-axis represents a different feature from the output query.
- The user can brush along the y-axis to filter the data in the table below the plot.
- Each new brush combines previously filtered values with the new data under brushing.

## Filtering options: Hovering



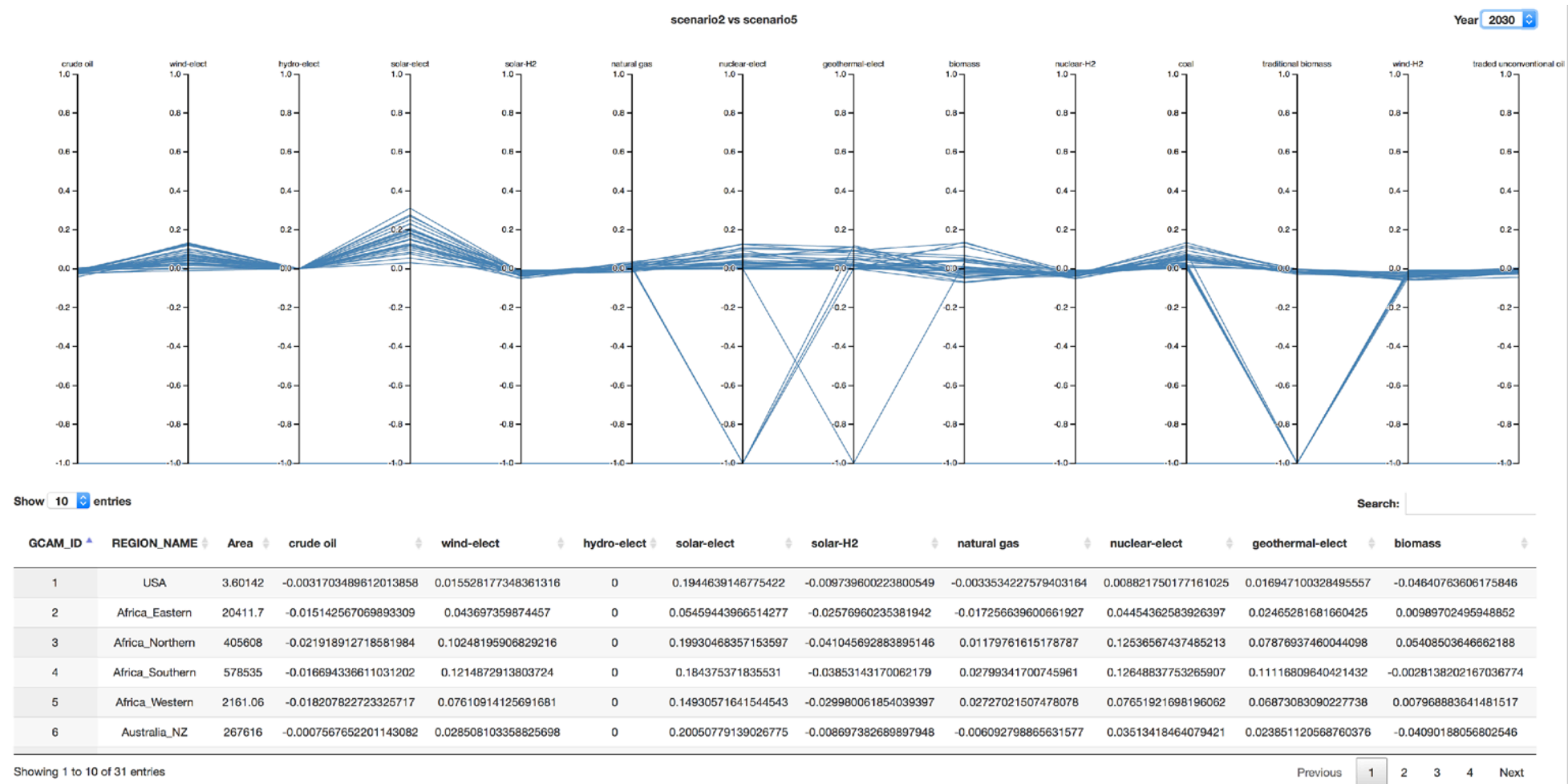
The user can hover over a line to examine a specific region or scenario in the data table below.

Show 10 entries

Search:

GCAM_ID ▲	REGION_NAME ▼	Area ▼	crude oil ▼	wind-elect ▼	hydro-elect ▼	solar-elect ▼	solar-H2 ▼	natural gas ▼	nuclear-elect ▼	geothermal-elect ▼	biomass ▼	nuclear-H2 ▼
16	European Free Trade Association	100547	0.06631058632062814	-0.8357920593854207	0	-0.787477406959368	0.15953154514544776	-0.027459327961076474	0	-0.025287826718558615	-0.03347175861434554	0.151784402996034

# Difference View (Scenario A vs Scenario B)



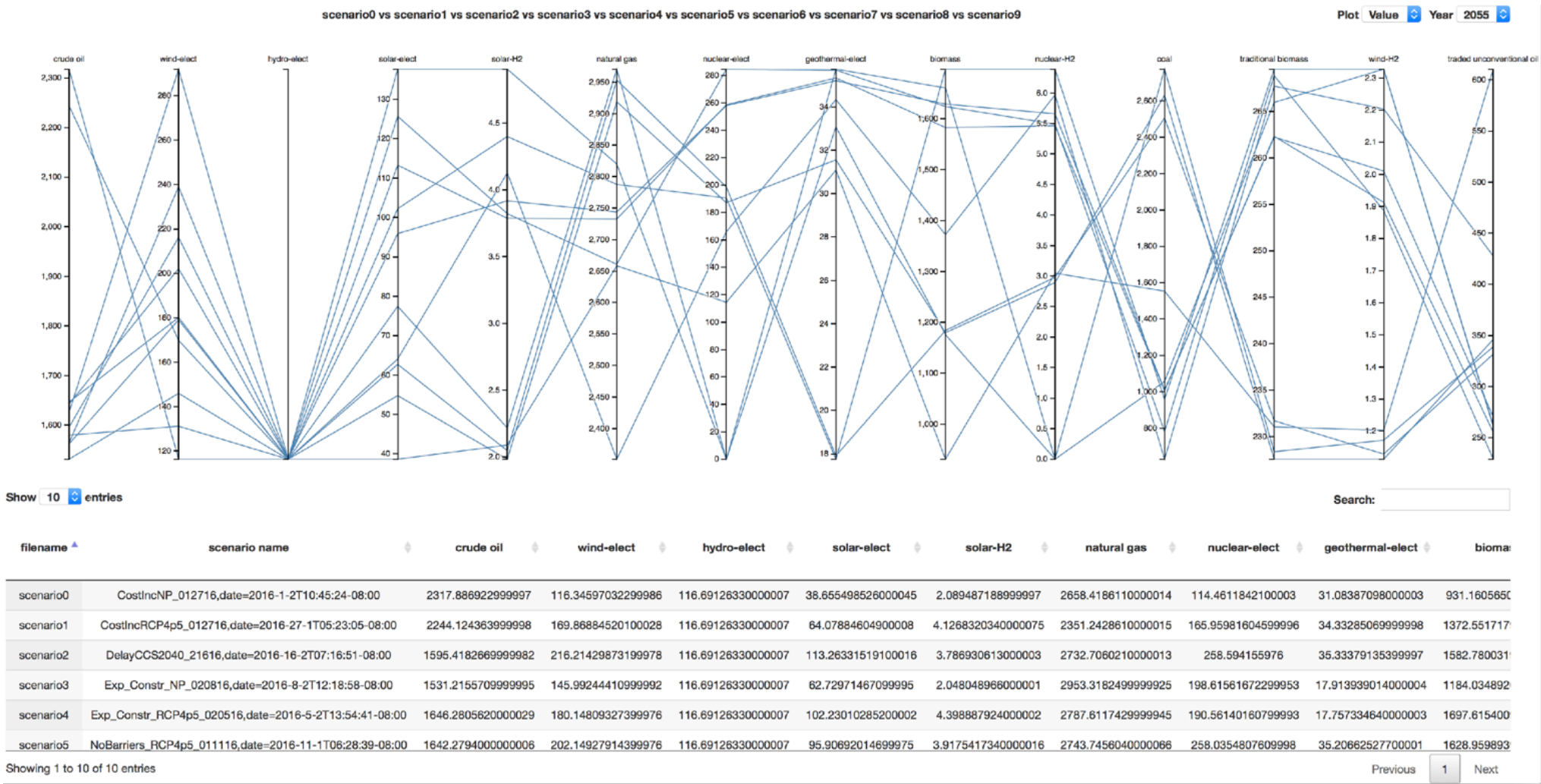
The Difference View shows the normalized difference of Scenario A vs Scenario B across all features.

## Difference View Cont.

- The normalized difference is found by rescaling each feature using feature scaling. Feature scaling is used to bring all values into the range [0,1]. The equation for feature scaling is defined as follows: 
$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$
- After rescaling each feature the difference of Scenario B is taken from Scenario A giving us a range of [-1,1].



Scenario Comparison View



The Scenario Comparison View shows the summation of all countries for each feature in the scenario files.

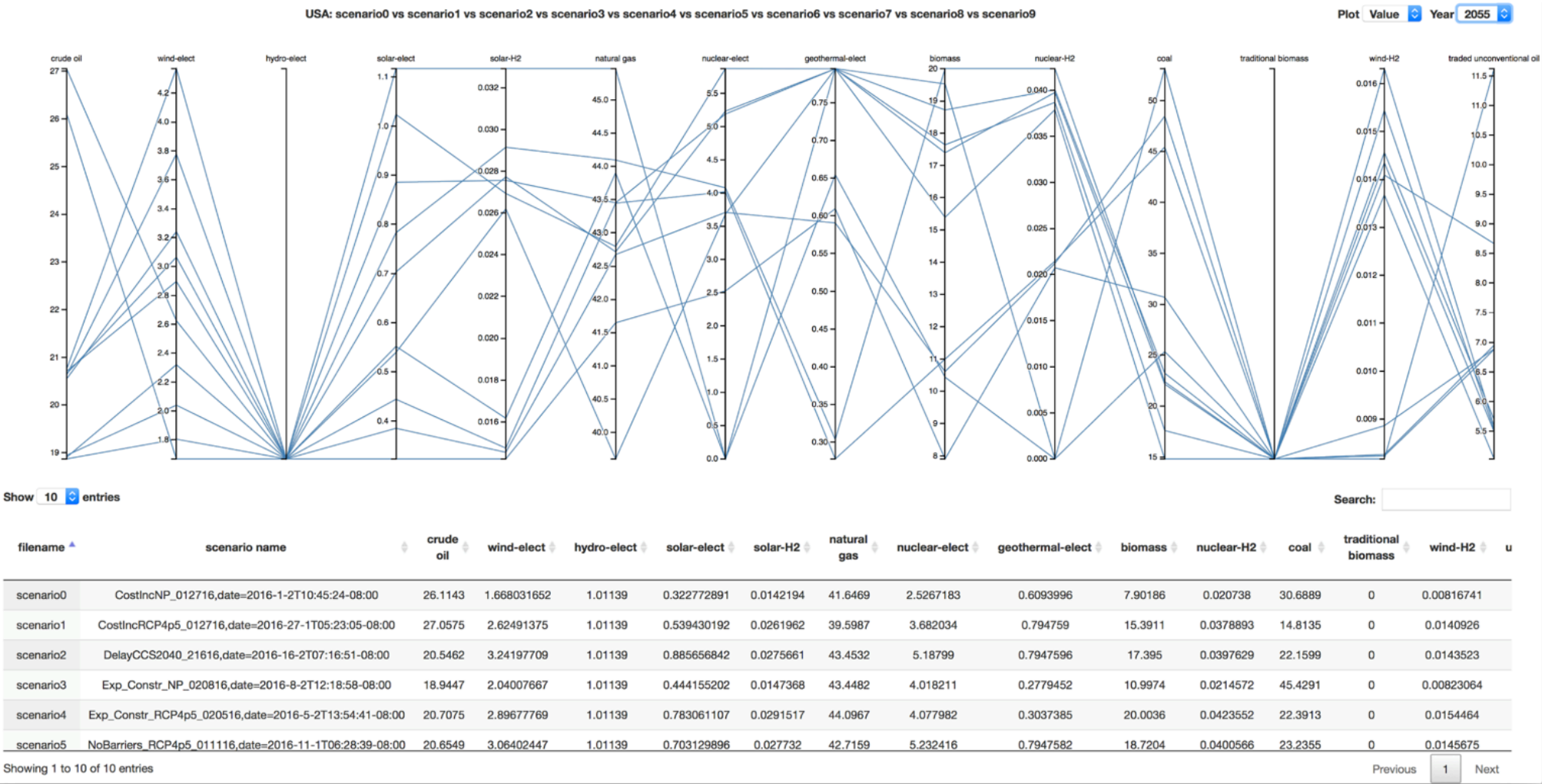


## Map View



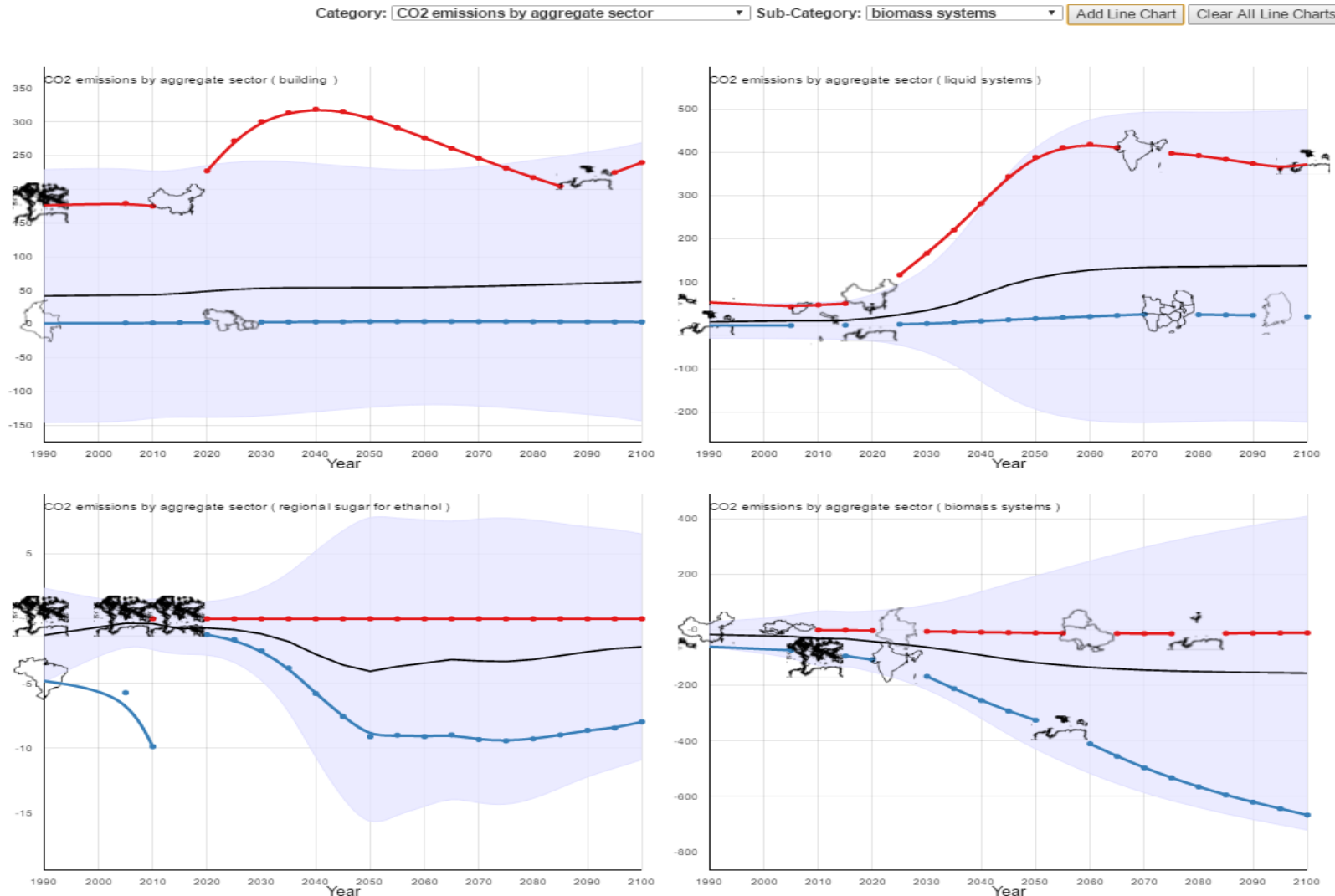
The Map View is used as a filter for the Parallel Coordinates View. When a region is selected the Parallel Coordinates View will be filtered to the selection.

# Region Comparison View



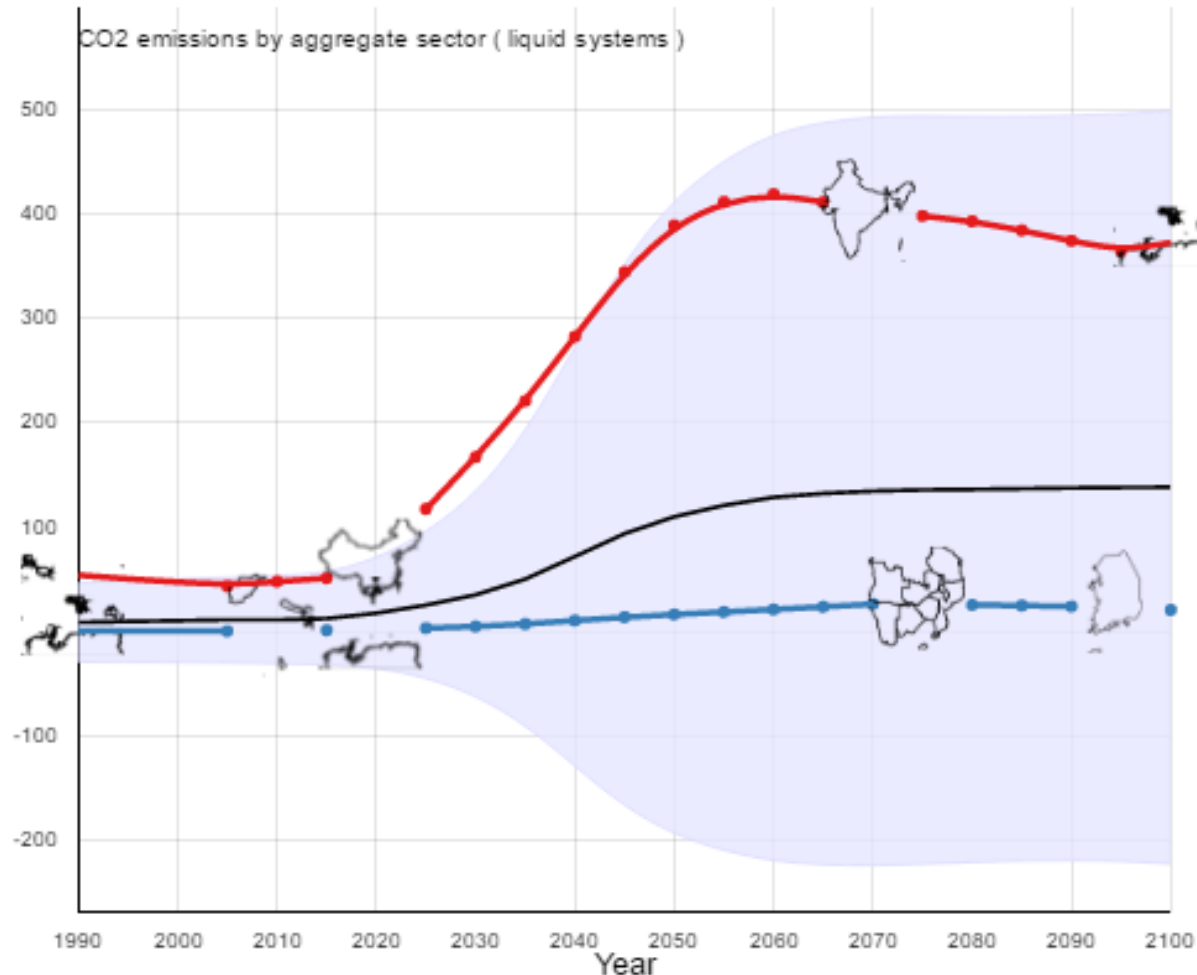
The Region Comparison View shows the value for the selected country across all scenario files.

# Min/Max View



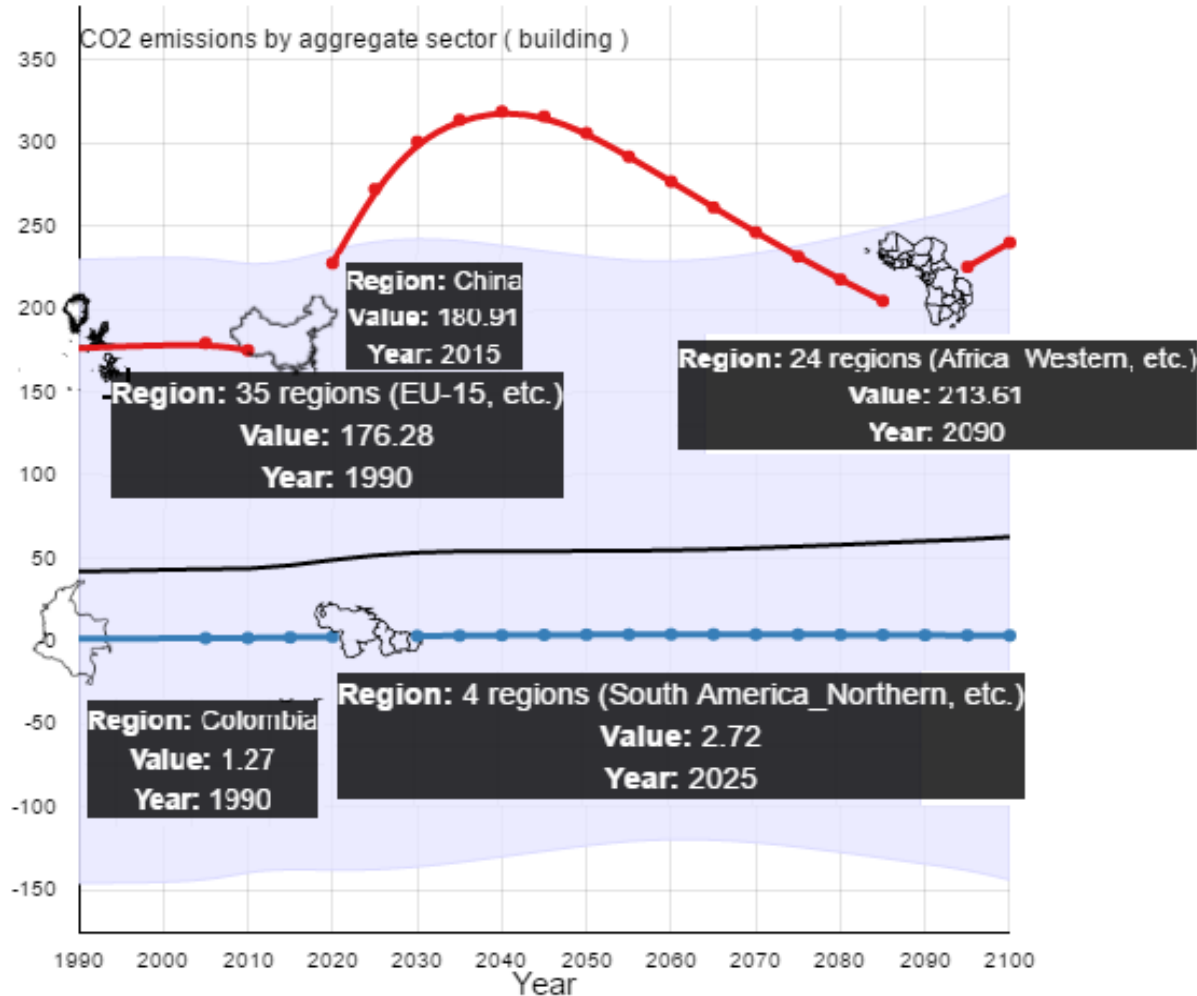
- Users can insert a min/max line chart for any of the output features from the provided queries.
- In the example on the left we have inserted line charts using CO2 emissions by aggregate sector for:
  - Building
  - Liquid Systems
  - Regional Sugar for Ethanol
  - Biomass Systems

## Min/Max View Cont.



- In each line chart, the x-axis represents the year of simulation and y-axis represents the output value.
- The black line represents the mean output value of all regions. The blue area chart represents the range mean  $\pm 2 * \text{std}$ .
- Each dot represents a set of regions ( or a single region) having the maximum or minimum value. The computation of output values for each country and each year is explained in the next slide.

## Min/Max View Cont.



- In the example on the left
  - The max value is shared by 35 regions and changes in 2015 to china. China continues to be the dominant region until 2090. During this period China soars above the 2 std area. Later in 2090 the max value is shared by 24 regions for the remainder of the simulation.
  - The min value is held by Columbia and this continues until 2025. From 2025 onto the end of the simulation 4 regions share the minimum value.

## Min/Max View Cont.

The analysis goal is to 1) find the abnormal countries in each type of output and each year from all scenarios, and 2) find similar outputs with similar changing trend.

Step 1: for each output at each country and each year, we compute its average value over all scenarios.

$$\overline{O_j}(C_i, t_k) = \sum_{p \text{ Scenario}(C_i, t_k)} / \# \text{ of Scenarios}$$

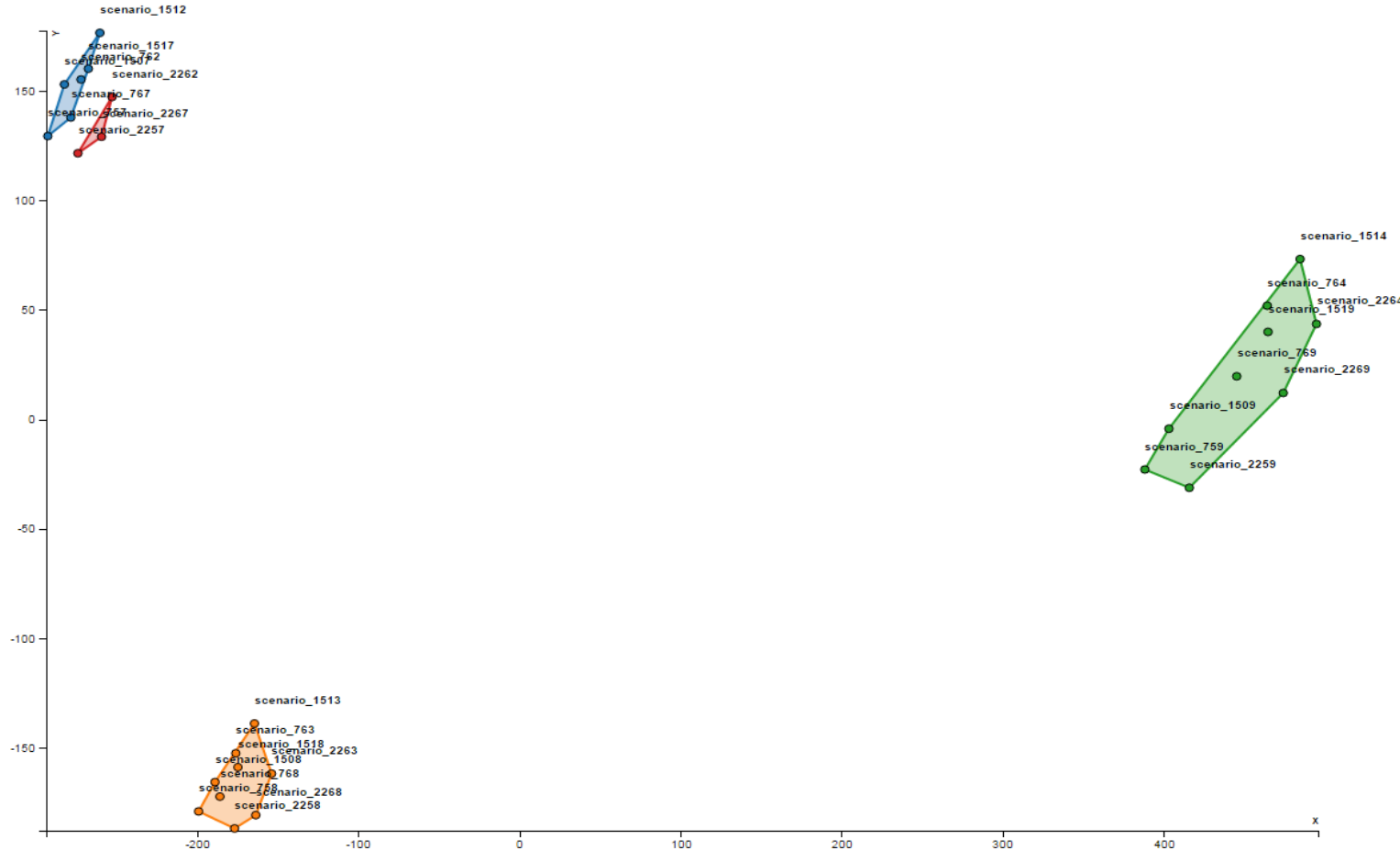
Step 2: So, for each output  $\overline{O_j}$  and each year  $t_k$ , suppose we have N countries ( $i=1 \dots N$ ), we can compute their mean and standard deviation over all countries:

$$\mu(\overline{O_j}) = \sum_{C_i} \overline{O_j} / \# \text{ of Countries}$$

$$\sigma(\overline{O_j}) = \sqrt{\sum_{C_i} (\mu(\overline{O_j}) - \overline{O_j})^2 / \# \text{ of Countries}}$$

Step 4: we can draw a line chart for each output of which the x-axis is the time  $t_k$  and y-axis is  $\mu(\overline{O_j})$ . We can also add some dots onto the line chart. The dots represent the abnormal countries having the maximal and minimal values.

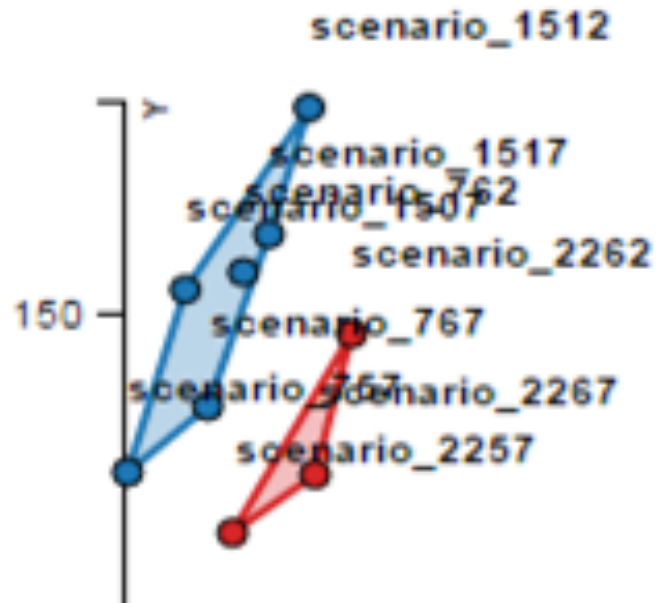
# Cluster View



- The Cluster View shows the PCA plot of all scenarios and their clusters from K-Means.
- This view provides a spatial overview of each scenario and shows their similarity based on locality and clustering.



## Cluster View Cont.



- Each dot represents a different scenario and the coloring is used to show which cluster they belong to.
- The User can control the number of clusters generated by K-Means.



Demo

## Thank You

- Contributions
  - Visual analytics tool for processing multiple scenarios
  - Tools for analyzing and comparing scenarios
- Future Work
  - More visualization features
  - UI improvements
  - Server distribution
  - Parallelization