Predicting International Carbon Dioxide and Greenhouse Gas Emission Levels with Dynamic Mode Decomposition

Ethan Biegeleisen
Manhattan College
MATG 511 - Computational Methods for Analytics
Fall 2019

1 Introduction and Background Information

The focus of this project is to utilize Dynamic Mode Decomposition (DMD) for two sections of the United Nations' International Greenhouse Gas (GHG) Inventory Data [7]. The first part of the data set is a collection of international carbon dioxide (CO2) emissions across countries and other areas from 1990 to 2014 [7]. The second part of the data set is a collection of international GHG emissions across the same countries and areas from 1990 to 2014 [7]. The project goals are to determine if DMD can provide accurate estimations of future CO2 and GHG emissions, and to compare the results from DMD on the CO2 and GHG parts of the data set. DMD could also potentially provide a linear regression of best fit to each part of the data set [3]. One major question is if DMD will result in similar types of estimations for both CO2 and GHG emissions or if it could give seemingly contradictory estimations for CO2 and GHG emissions. An example of this would be if CO2 emissions are estimated to decrease but GHG emissions are estimated to increase.

The importance of researching greenhouse gas emissions is that they are the main culprit behind climate change [2]. The global mean surface air temperature is estimated to have increased by 1.53 degrees Fahrenheit from 1880 to 2012, and estimates for global average temperature in future decades depend on the amount of future global greenhouse gas emissions [2]. Major ramifications of increased global average temperature include increased risks of floods and droughts, increased sea level, stronger hurricanes, and the extinction of numerous species [2]. CO2 is the heaviest contributor to greenhouse gas emissions, with GHG also including gases like methane, ozone, water vapor, chlorofluorocarbons, and nitrous oxide [2]. The curbing of emissions for GHG and CO2 in particular are pivotal in minimizing the effects of climate change.

One group that conducted successful predictions based on a data set using a variation of DMD are Clainche, Lorente, and Vega [1]. They collected data on the upstream velocity of a wind turbine from six different locations over a 24 hour period using light detection and ranging measurements [1]. Over the same 24 hour period, Higher Order Dynamic Mode Decomposition (HODMD) was utilized to successfully predict the wind turbine's upstream velocity at a position outside the range of the light detection and ranging measurements [1]. Clainche et al. were able to apply a variation of DMD on a data set to correctly approximate a physical occurrence in the real world. Additional work using DMD comes from Proctor and Eckhoff, who utilized it to map out data sets for Google's trends for indicating levels of flu activity in the United States, mapping out cases of pre-vaccinated measles in the United Kingdom, and mapping out cases of type 1 polio in Nigeria [5]. Their results show

how DMD can be used to organize and map out data sets when sections of a data set originating from different regions. Meanwhile, Mann and Kutz found that DMD has potential for successful use in stock market trends if it works with a learning algorithm to improve sampling and prediction windows [3]. This presents another way in which DMD can be utilized to provide estimates of data in near-future time frames. Mohan, Soman, and Sachin Kumar found another successful use of DMD for short-term estimates by predicting energy demand for electricity a few hours to weeks in advance [4].

2 Methods

DMD is a mathematical tool that relies on measurements from different points in time like historical data, and the tool is generally considered to be equation-free [5]. This means DMD does not attempt to find a set of equations that directly calculate the entire data set [5]. Instead, DMD approximates a nonlinear set of data into a linear model. This linear model is a least-square fit system that is able to provide estimations for near-future states of the data set [3]. In addition, DMD is useful because its process includes dimension reduction [3]. DMD can be utilized for reconstruction as well, such as recovering data with gaps in their measurements or signals [6]. If the United Nations data set is split into one part for the CO2 emissions and one part for the GHG emissions, the CO2 part should be able to go through DMD to provide a least-square fit system that estimates the amount of CO2 emissions per year in the near-future and the GHG part should be provided its own least-square fit system through DMD that estimates the near-future amount of GHG emissions per year.

To utilize DMD, each part of the data set first needs to be split into two matrices. As an example, this section will assume DMD is being performed on the part of the data set that contains all of the information on CO2 emissions. The two matrices are:

$$X = [\bar{x}_1, \bar{x}_2, \bar{x}_3, ...\bar{x}_{m-1}] \tag{1}$$

$$\bar{X} = [\bar{x}_2, \bar{x}_3, \bar{x}_4, ...\bar{x}_m] \tag{2}$$

X is an array of vectors in which $\bar{x}_1$ is a column containing the CO2 emissions of every country or area in the year 1990 and $\bar{x}_{m-1}$ is is a column containing the CO2 emissions of every country or area in the year 2013. Similarly, $\bar{X}$ is an array of vectors in which $\bar{x}_2$ is a column containing the CO2 emissions of every country or area in the year 1991 and $\bar{x}_m$ is a column containing the CO2 emissions of every country or area in the year 2014. Essentially, X is a matrix containing the CO2 emissions for every country or area each year from 1990 to 2013 and $\bar{X}$ is a matrix containing the CO2 emissions for every country or area each year from 1991 to 2014.

The next step is to find the rank of X and let the rank be equal to r, such that r is equal to the number of linearly independent columns in X. Singular Value Decomposition (SVD) is performed next with rank truncation, which creates a reduced square matrix, X_r. The number of rows and columns X_r contains is equal to the rank of X such that:

$$X_r = U_r \Sigma_r V_r^* \tag{3}$$

U is an array of the eigenvectors of XX^*, V is an array of the eigenvectors of X^*X, and Σ is a rectangular diagonal matrix whose diagonal entries are called σ. U_r is a matrix containing the first r columns of U, V_r is a matrix containing the first r columns of V, and Σ_r is a square matrix containing the first r rows and the first r columns of Σ. V_r^* is the conjugate transpose of V_r. The square matrix A_r is calculated afterwards:

$$A_r = U_r^* \bar{X} X^+ U_r \tag{4}$$

$$A_r = U_r^* \bar{X} V_r \Sigma_r^{-1} \tag{5}$$

X^+ is the pseudo-inverse of X. A non square matrix like X does not have a standard inverse, so the pseudo-inverse is considered the closest possible solution to finding the inverse of a non square matrix [3]. The next step is to find the eigenvectors and eigenvalues of A_r such that W is an array in which each column is an eigenvector of A_r and $[\lambda_1, \lambda_2, ..., \lambda_r]$ is an array in with each entry is an eigenvalue of A_r. After this, the following is computed:

$$\Phi = \bar{X} V_r \Sigma_r^{-1} W \tag{6}$$

Φ is a matrix whose columns are called the DMD modes. The next step is to calculate one of the two following equations. One option is the discrete method, which is the following:

$$x_k = \sum_{j=1}^{r} \phi_j \lambda_j^{k-1} b_j \tag{7}$$

The other option is the continuous method, which is the following:

$$\vec{x}(t) = \sum_{j=1}^{r} \phi_j e^{\omega_j t} b_j \tag{8}$$

x_k or $\vec{x}(t)$ is a column vector whose entries are the CO2 emissions of each country or area for that year and ω_j is equivalent to $\frac{log(\lambda_j)}{\Delta t}$. Δt is equivalent to $\frac{1}{m-1}$ in which m is equivalent to the range of t values from t_1 to t_m. b_j is a column from a matrix called B, and B can be solved for with the following equation:

$$B = \Phi^+ \vec{x}(0) \tag{9}$$

The pseudo-inverse of Φ is used because Φ is not always a square matrix [3]. Once these last equations have been calculated, the amount of CO2 emissions after the years in original data set can be estimated using equation (7) or equation (8), which are both linear models. If $\vec{x}_0$ in equation (7) or $\vec{x}(0)$ in equation (8) is a column whose entries are the CO2 emissions of each country or area in 1990, then $\vec{x}_{30}$ in equation (7) or $\vec{x}(30)$ in equation (8) gives a column whose entries are approximations of the amount of CO2 emissions from each country or area in 2020. To use DMD for the GHG data, equations (1) through (9) are repeated using the GHG data as a basis for equation (1) and equation (2) instead of the CO2 data.

The United Nations data set is sorted into four columns. The first column provides the name of the country or region the CO2/GHG emission value is from, the second column provides the year the CO2/GHG emission is from, the third column provides the CO2/GHG emission value, and the fourth column states which category of emission the value is from. The emissions are presented in reverse chronological order by year for each country or area, and the countries or areas are presented in alphabetical order. For example, the first row is for Australia in 2014, the second row is for Australia in 2013, and this follows until reaching the last row for Australia in 1990. The following row is for Austria in 2014 and the process repeats until the CO2 emission values for each country in each year has been provided. Then this entire process repeats for the GHG values of each country in each year. The first section of the data set provides all the CO2 emission values and the second section of the data set provides all of the GHG emission values, but there are also sections below the GHG emission values for other gases like methane and nitrous oxide. The CO2 emission values are measured in kilotons without land use, land-use change, and forestry. The GHG emission values are also measured in kilotons and include CO2 emission values without land use, land-use change, and forestry. Figure 1 displays all of the data set's CO2 emission values and GHG emission values from Australia.

The countries and areas included in the United Nations data set that contain 1990-2014 data on CO2 emissions and GHG emissions are the following: Australia (1), Austria (2), Belarus (3), Belgium (4), Canada (5), Croatia (6), Cyprus (7), the Czech Republic (8), Denmark (9), Estonia (10), the European Union (11), Finland (12), France (13), Germany (14), Greece (15), Hungary (16), Iceland (17), Ireland (18), Italy (19), Japan (20), Latvia (21), Lithuania (22), Malta (23), the Netherlands (24), New Zealand (25), Norway (26), Poland (27), Portugal (28), Romania (29), the Russian Federation (30), Slovakia (31), Slovenia (32), Sweden (33), Switzerland (34), Turkey (35), Ukraine (36), and the United States (37). The number next to each location corresponds to where the location can be found in the figures for the Results section. Additional countries and areas in the data set are Bulgaria, Liechtenstein, Luxembourg, Monaco, Spain, and the United Kingdom. However, Monaco's CO2 data covers 1990-2013 instead of 1990-2014 so it needs to be excluded from the data set in order to accurately utilize DMD. Bulgaria, Liechtenstein, Luxembourg, Spain, and the United Kingdom also need to be excluded because they include CO2 data, but not the corresponding GHG data.

To prepare the data set for DMD, a Numerical Python (NumPy) array is created from the CO2 emission values for each country or area from 2014 to 1990. The array is then flipped so the values go from the year 1990 to the year 2014. Once each country or area has its own NumPy array, the above matrix X is created using the CO2 values for each country or area from 1990 to 2013 and the above matrix $\bar{X}$ is created using the CO2 values for each country or area from 1991 to 2014. This process is repeated for the GHG emission values for each country or area except the resulting matrices are called Y and $\bar{Y}$ instead. Once X, $\bar{X}$, Y, and $\bar{Y}$ are created, the data is ready for DMD. Table 1 displays a small section of X from equation(1) with labels added for each location and year.

Location / Year	1990	1991	1992	...	2013
Australia	278266	279742	284766	...	396914
Austria	62297	65904	60432	...	67957
Belarus	100438	93649	87061	...	61511
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
United States	5115095	5064880	5170274	...	5502551

Table 1: Annual CO2 emissions emissions in each location (rounded to the nearest integer). Data set from [7].

3 Analysis of Results

Discrete and continuous DMD were both conducted successfully on the CO2 emissions data set and the GHG emissions data set. The first comparison that was attempted from the DMD models was between the discrete DMD estimate for CO2 emissions in 2010, the continuous DMD estimate for CO2 emissions in 2010, and the data set's records of CO2 emissions in 2010. Figure 2 displays this comparison. The numbers on the x-axis of each figure in this section match the numbers given in the previous section to each country or area in the data set. For example, Australia corresponds to 1 on the x-axis, Austria corresponds to 2 on the x-axis, and the United States corresponds to 37 on the x-axis.

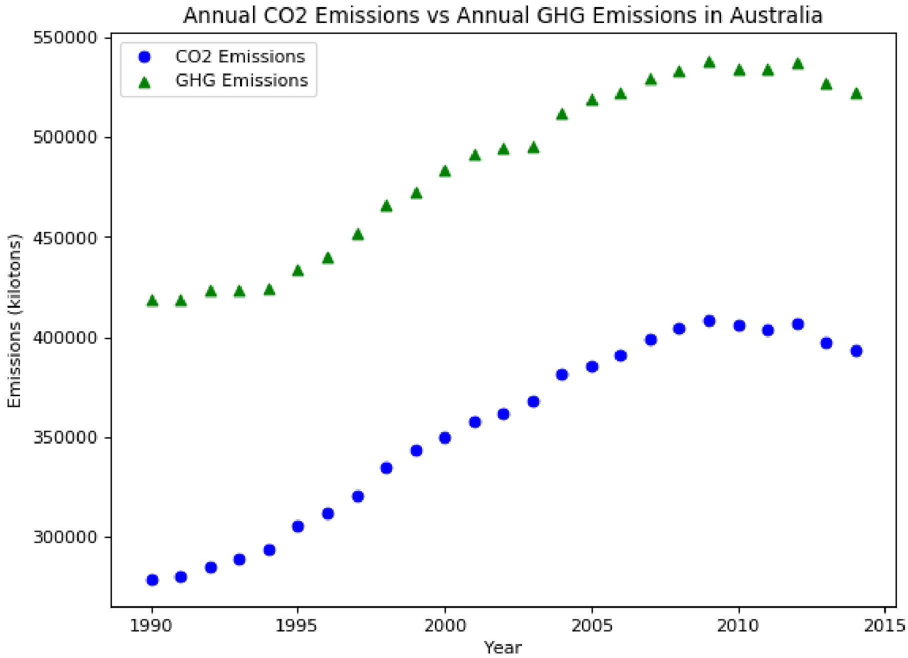

Figure 1: Annual CO2 emission values from Australia compared to annual GHG emission values from Australia. Data set from [7].

It should be noted that all of the DMD models for both CO2 emissions and GHG emissions resulted in entries with complex values. When plotting these data points, Python only plots the real number in the entry and ignores the imaginary number. As a simplified example, if one of the entries is $50 + 3i$, then Python only plots the number 50. However, the loss of the imaginary values in the plots does not seem to have a significant impact on the plots because the real values in each entry are many times larger than the imaginary values in each entry.

Figure 2 shows that the discrete and continuous DMD models gave the same results for CO2 estimates in 2010. In addition, both DMD models gave CO2 emissions estimates that were close in accuracy to the large majority of countries or areas included in the data set. The three major exceptions were the European Union, the Russian Federation, and the United States which all had drastically higher records of CO2 emissions than the DMD models expected. However, the DMD models still seemed to mostly capture the general trend of CO2 emissions in each country or area.

The DMD models for GHG estimates in 2010 generally displayed similar types of results to the CO2 estimates with one major exception. Figure 3 shows that the discrete and continuous DMD models gave the same GHG estimates in 2010. However, the DMD models for GHG were much more accurate at predicting 2010 GHG emissions from the European Union, the Russian Federation, and the United States than the DMD models for CO2 were at predicting the 2010 CO2 emissions from the same three locations. Overall, the DMD models for GHG emissions appear to be highly accurate.

Due to both CO2 DMD models finding the same results in Figure 2 and both GHG models finding the same results in Figure 3, the rest of the results in this section will come from the discrete DMD CO2 model and the discrete DMD GHG model. Figure 4 displays predicted CO2 emissions from 2015 to 2020 using the discrete DMD model. 2015 is the first year included in the figure because the

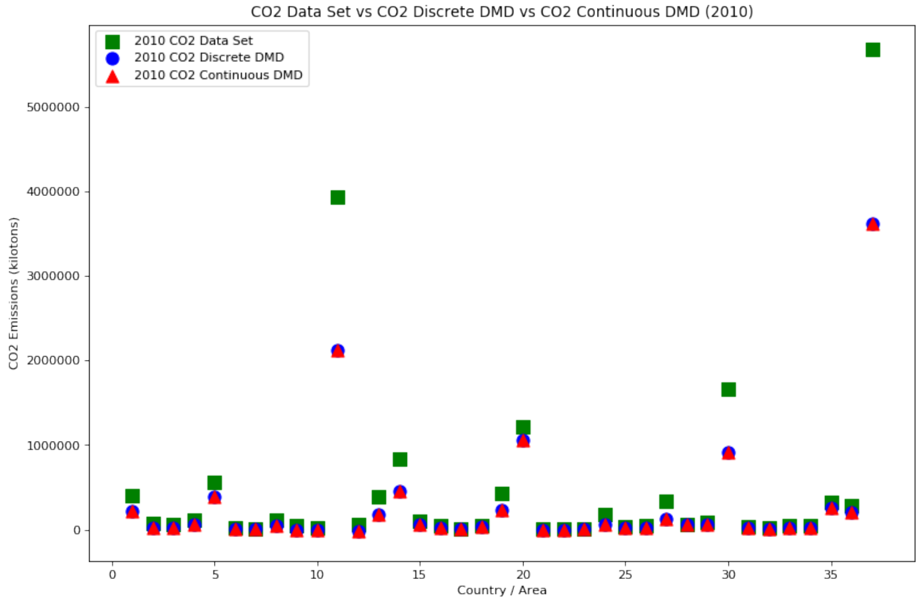

Figure 2: Comparing 2010 CO2 emission values from the UN data set [7] with the 2010 estimates from the discrete DMD CO2 model and the 2010 estimates from the continuous DMD CO2 model. Each number on the x-axis corresponds to one of the countries or areas in the data set.

original data set ended at 2014. Most of the locations are predicted to have roughly the same level of CO2 emissions or slight increases in emissions from 2015 to 2020. The general trends are similar to the ones in Figure 2 except the outliers with higher annual CO2 emissions like the European Union, Japan, the Russian Federation, and the United States show much more variance in the year-to-year predictions than the locations with lower annual CO2 emissions. By 2019 and 2020, the DMD model also starts predicting some locations to have a negative annual CO2 emissions value.

Figure 5 displays predicted GHG emissions from 2015 to 2020 using the discrete DMD model, and its general trends are similar to the trends in Figure 3. The Figure 5 results seem mostly similar to the Figure 4 results, except with much less variation in the outlier locations than Figure 4 has. Most of the locations are still predicted to have roughly the same emission levels or slightly higher emission levels from 2015 to 2020. In addition, none of the discrete DMD estimates in Figure 5 seem to predict a negative GHG emission value from any of the countries or areas in 2019 or 2020.

In Figure 6, the discrete DMD model's predictions for CO2 emissions in 2015 and 2016 are directly compared with reports on the CO2 emissions for each location in 2015 and 2016 [8]. Similar to Figure 2, the DMD model appears to match the reported data fairly well aside from the major emissions outliers like the European Union and the United States.

Figure 7 compares the predicted GHG emissions in 2015 and 2016 from the discrete DMD model with reported GHG emissions from each location in 2015 and 2016 [9]. However, one additional note needs to be made about the data used in this comparison. The DMD models for GHG emissions are based on a data set of GHG emissions that include indirect CO2, so the DMD models for GHG emissions are meant to predict GHG emissions including indirect CO2. However, the 2015 and 2016 GHG data used in Figure 7 to compare against the DMD predictions does not appear to include indirect CO2. As a result, the comparison in Figure 7 might have more limited findings than the

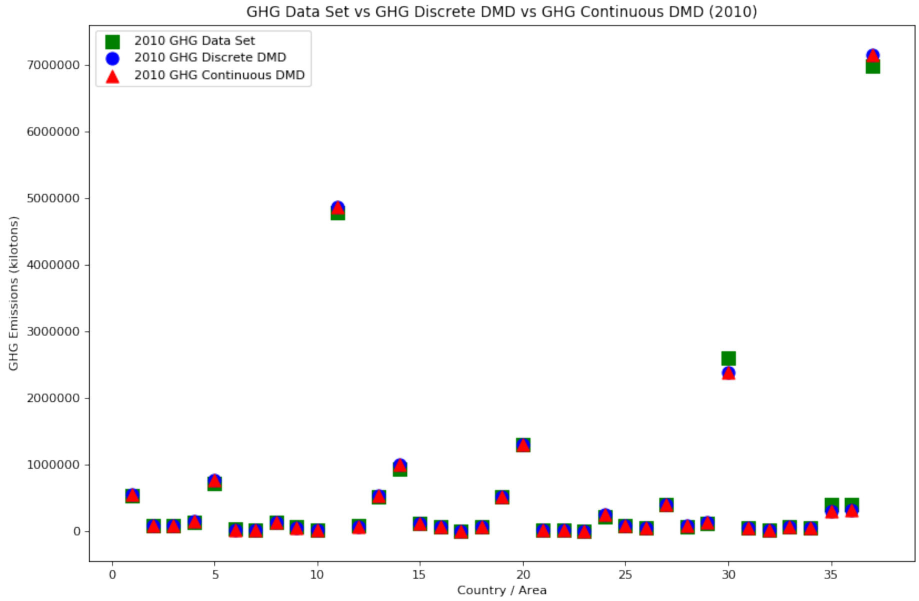

Figure 3: Comparing 2010 GHG emission values from the UN data set [7] with the 2010 estimates from the discrete DMD GHG model and the 2010 estimates from the continuous DMD GHG model. Each number on the x-axis corresponds to one of the countries or areas in the data set.

comparison in Figure 6. However, the DMD predictions for GHG emissions in 2015 and 2016 appear to match the recorded data on GHG emissions in 2015 and 2016 quite closely. Once again, the locations with much higher CO2 and GHG emissions show less variance in the GHG predictions than in the CO2 predictions.

Figures 8 and 9 compare the accuracy of discrete DMD reconstruction for Australia's CO2 emission levels and GHG emission levels. Figure 8 shows that while the general trend for Australia's CO2 emission levels appears to be decently reconstructed by the discrete CO2 DMD model, the model's reconstruction provides significantly lower values for Australia's CO2 emission levels than the original data set. In comparison, the discrete GHG DMD model provides a much more accurate reconstruction of Australia's GHG emission levels.

A few trends are visible from the results of this project. One consistency is that the DMD models for CO2 emissions seem to give similar types of results as the DMD models for GHG emissions. None of the figures show any of the CO2 DMD models having opposite trending predictions compared to the GHG DMD models. Another consistency is that the DMD models for GHG appear to consistently offer more accurate reconstructions and short-term predictions on future emissions than the DMD models for CO2. In addition, the GHG DMD models show less variance in outlier locations like the European Union, the Russian Federation, and the United States than the CO2 DMD models. The discrete GHG DMD model also shows less variance in 2015-2020 predictions than the discrete CO2 DMD model. Although the GHG DMD model does start to show some variance in outlier locations as later years are predicted in Figure 5, this variance is substantially smaller than the amount of variance shown in Figure 4 among 2019 and 2020 predictions for the CO2 DMD model. The CO2 DMD model starting to display negative values in 2019 and 2020 might also signify that the model's short-term use is nearing its end by the time it reaches those years. One other egregious

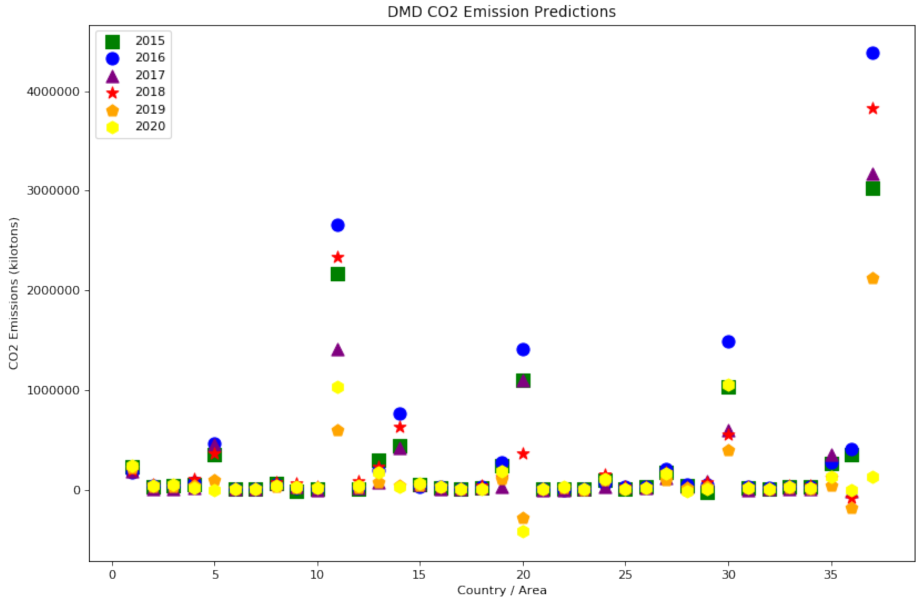

Figure 4: Predicted 2015-2020 CO2 emissions using the discrete DMD model for CO2 emissions. Each number on the x-axis corresponds to one of the countries or areas in the data set.

example of this in Figure 4 is that the CO2 DMD model predicted the United States to have a similar level of CO2 emissions as Poland and Turkey in 2020.

A major question that stems from these findings is why the GHG DMD model would consistently present more accurate results the CO2 DMD model when the methods used to create the DMD models were identical. The matrices X and Y each have a rank of 24 so one possibility might be that X's rank is too large to provide more accurate predictions through DMD. However, that would not address why the GHG DMD model is consistently more accurate in spite of Y having the same rank as X. All of the differences between the CO2 DMD model and the GHG DMD model should stem from the DMD models being based on different data sets, which could imply the entries in the CO2 data set are leading to more variance in the CO2 DMD model than the entries in the GHG data set are for the GHG DMD model. If this was the case, then access to 50 years of emissions for both data sets instead of 25 years of emissions might lead to more consistent amounts of variance between the CO2 DMD model and the GHG DMD model. The DMD models might also present more similar results if the data on emission levels for many more countries and areas were available. However, the predictions from the CO2 and GHG DMD models still appear to give relatively consistent trends in emission levels.

4 Discussion

Utilizing DMD for this project ended up being a bit more straightforward than expected. Although DMD must be conducted in a specific order with quite a few stages involved, each individual stage was not too difficult to program. Focusing on each stage one step at a time while programming might have made working with DMD feel less intimidating. Permission to use inbuilt Python functions also helped like numpy.linalg.svd for conducting SVD and numpy.linalg.pinv for finding

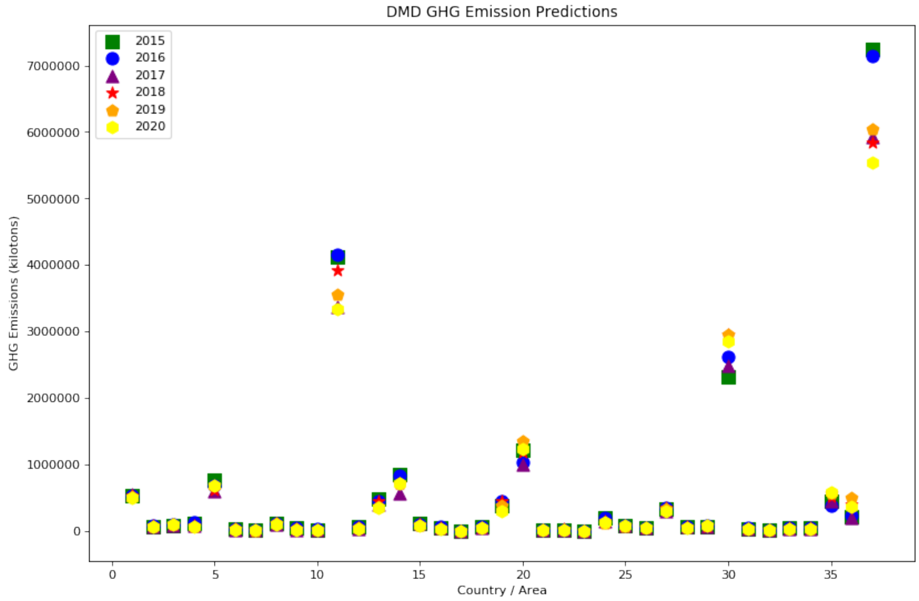

Figure 5: Predicted 2015-2020 GHG emissions using the discrete DMD model for GHG emissions. Each number on the x-axis corresponds to one of the countries or areas in the data set.

the pseudo-inverse of a matrix. However, properly preparing the data set for DMD to create X, $\bar{X}$, Y, and $\bar{Y}$ took a bit more time than expected.

DMD is a relatively new mathematical tool, so there appear to be numerous potential applications for it that have not yet been utilized. While conducting research for this project, no studies were found that had used DMD to predict CO2 or GHG emissions. This could potentially be the first work utilizing DMD to predict CO2 emissions, to predict GHG emissions, and to compare the results of those DMD predictions using a data set of CO2 and GHG emission levels from countries and areas around the world.

For this project, a data set containing annual CO2 and GHG emission levels in 37 countries or regions from 1990 to 2014 was used as a basis for discrete and continuous DMD. The goals were to determine if DMD could be utilized to accurately predict future CO2 and GHG emissions, as well as comparing the CO2 predictions with the GHG predictions to examine how consistent the predictions were between the two emission groupings. The CO2 DMD models appeared to accurately predict the near-future trends in CO2 emission levels, but the models had more difficulty accurately reconstructing and predicting the specific emission levels each year. The CO2 DMD models had particular trouble predicting near-future emission levels from locations with a history of much higher CO2 emission levels than the rest of the locations in the data set like the European Union and the United States. However, the GHG DMD models were able to more accurately predict near-future GHG emission levels from these outlier locations in addition to accurately predicting near-future GHG emission levels for the rest of the locations in the data set. The GHG DMD models also provided a more accurate reconstruction of GHG emission levels overall. The general prediction trends were also consistent between the CO2 and GHG DMD models overall.

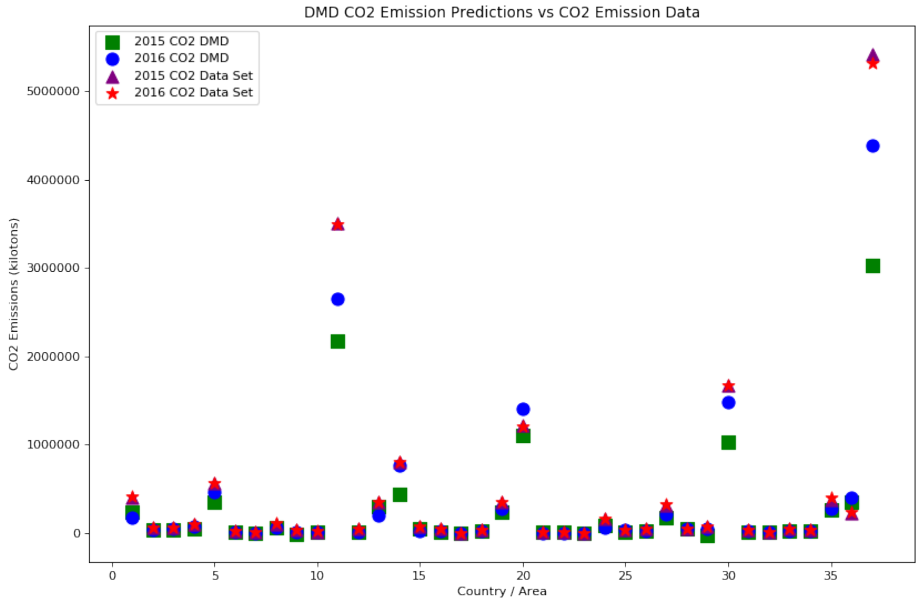

Figure 6: Comparing predicted 2015 and 2016 CO2 emissions from the discrete DMD model for CO2 emissions with reported 2015 and 2016 CO2 emission values. Data set from [8].

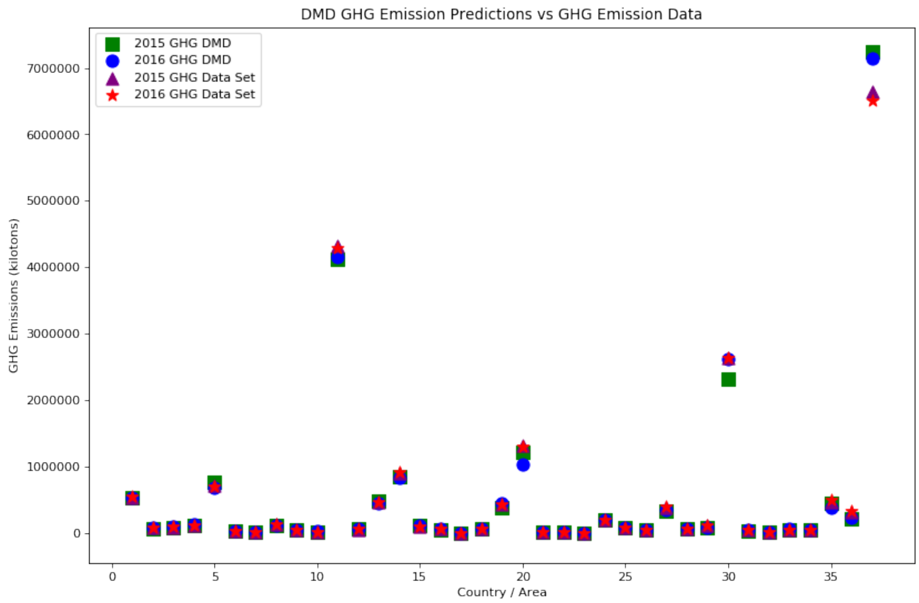

Figure 7: Comparing predicted 2015 and 2016 GHG emissions from the discrete DMD model for GHG emissions with reported 2015 and 2016 GHG emission values. Data set from [9]. Note: The DMD model predicts for GHG emission values including indirect CO2, while the 2015 and 2016 data set values do not appear to include indirect CO2.

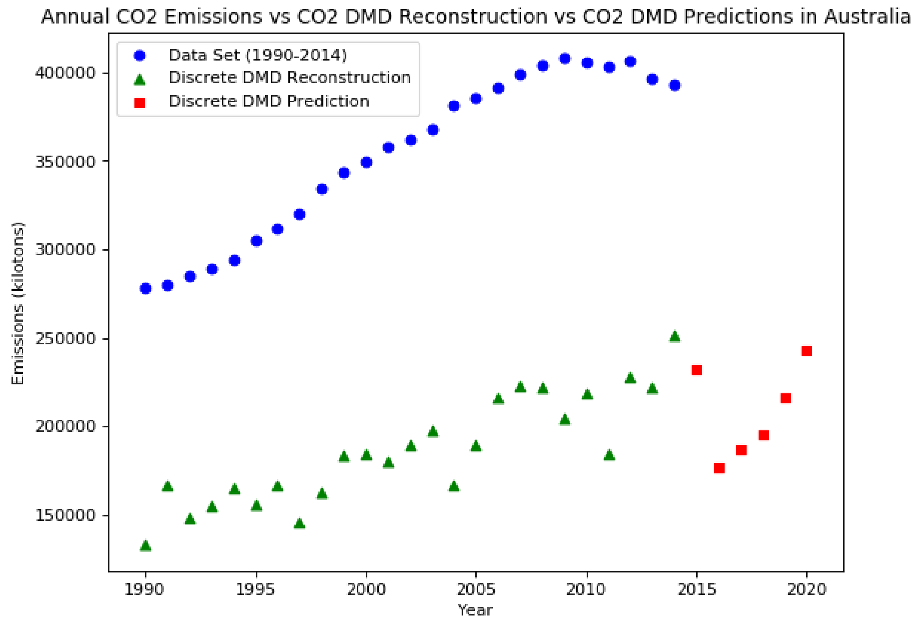

Figure 8: Comparing the discrete DMD model for Australia's CO2 emission levels with Australia's documented CO2 emission levels. Data set from [7].

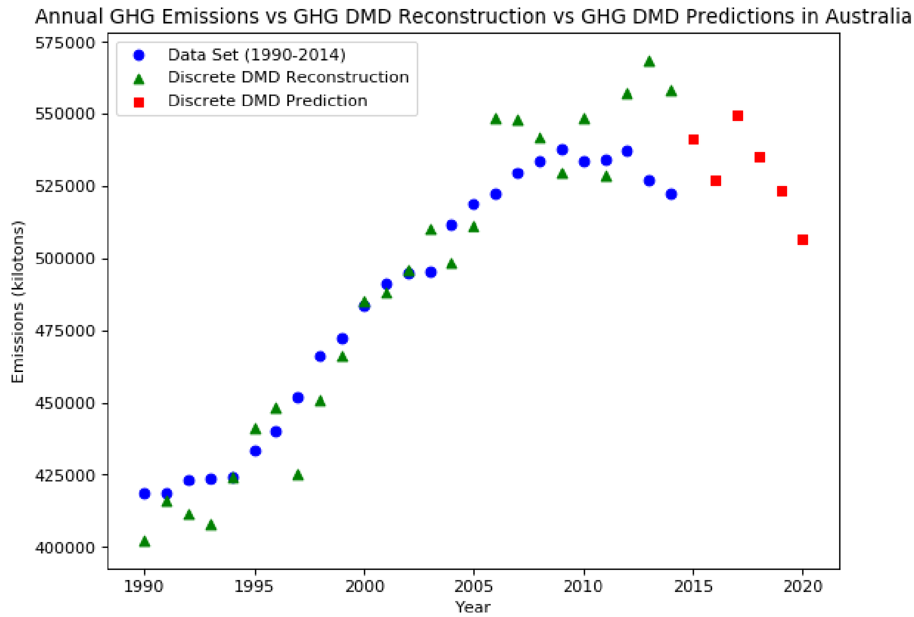

Figure 9: Comparing the discrete DMD model for Australia's GHG emission levels with Australia's documented GHG emission levels. Data set from [7].

References

[1] Clainche, S.L., Lorente, L.S., & Vega, J.M. (2018). Wind predictions upstream wind turbines from a LiDAR database. *Energies, 11*(4), 543. Retrieved from: `https://www.mdpi.com/1996-1073/11/3/543`

[2] Henson, R. (2014). *The thinking person's guide to climate change*. Boston: American Meteorological Society.

[3] Mann, J., & Kutz, J.N. (2015). Dynamic mode decomposition for financial trading strategies. *Quantitative Finance, 16*(11). doi: 10.1080/14697688.2016.1170194

[4] Mohan, N., Soman, K.P., & Sachin Kumar, S. (2018). A data-driven strategy for short-term electric load forecasting using dynamic mode decomposition model. *Applied Energy, 232*, 229-244. Retrieved from: `https://doi.org/10.1016/j.apenergy.2018.09.190`

[5] Proctor, J.L., & Eckhoff, P.A. (2015). Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International Health, 7*(2), 139-145. doi: 10.1093/inthealth/ihv009

[6] Schmid, P.J. (2011). Application of the dynamic mode decomposition to experimental data. *Experiments in Fluids, 50*(4), 1123-1130. doi: 10.1007/s00348-010-0911-3

[7] United Nations (2017). International Greenhouse Gas Emissions: A global GHG inventory from 1990-2017. Retrieved from: `https://www.kaggle.com/unitednations/international-greenhouse-gas-emissions`

[8] United Nations (2019). Carbon dioxide (CO2) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in kilotonne CO2 equivalent. Retrieved from: `http://data.un.org/Explorer.aspx`

[9] United Nations (2019). Greenhouse Gas (GHGs) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in kilotonne CO2 equivalent. Retrieved from: `http://data.un.org/Explorer.aspx`