A Causal Analysis on Time Series Data in the Earth System

Arnav Khanna
arkhanna@ucsd.edu

Dean Carrion
dcarrion@ucsd.edu

Keenan Hom
kwhom@ucsd.edu

Nithilan Muruganandham
nmuruganandham@ucsd.edu

Biwei Huang
bih007@ucsd.edu

Babak Salimi
bsalimi@ucsd.edu

Abstract

The goal of this study is to find causal relations in time series data within the earth system, as an alternative to demanding physical simulation models which are the current norm. This will be accomplished through the analysis of several variables with a focus on man-made ones, namely global temperature, gas emissions, electricity generation, and more. One of the methods used to detect causal relation is through neural networks, which looks for time-delayed relations. The second method used was Structural Causal Models, where we first learned instantaneous causal relations within the data, and then applied an Additive Noise Model to search for directed relations. We are also able to combine the two methods, to search for both instantaneous and time-delayed causal relations. Through these three methods, we were able to create three different causal models and present causal relations between the features within our data.

1 Introduction

Climate change is one of the most researched topics in the scientific community. The importance of studying such a topic cannot be understated, as earth system studies help us understand other aspects of science such as extreme weather, air quality, and human health. To better understand the driving forces behind the massive system that is earth is important, not just for planet longevity, but also for human longevity. From previous research, we understand that various human activities are the main contributing factors towards negative climate change. Understanding which of these activities are the most impactful helps us target and stop or slow the systems in place that harm our environment, protecting the earth and its inhabitants.

One way we can establish an understanding of earth system science is through studying which factors are driving forces behind other phenomena. For example, we would want to find out what drives global warming or air quality decreasing, so that we may try to stop or mitigate the source of these. One way to understand driving forces is through correlation, however, the caveat is that establishing correlation in no way establishes an undeniable link between two factors. Two events can be correlated in trends but unrelated in every other aspect, such as when two variables are caused be a third, unseen variable. This is why we must establish causation instead. Establishing causation is a different, more difficult method that does establish a link between two variables. These can be done through a number of causal methods, such as Granger causality and Nonlinear state-space methods.

As mentioned in Song and Ma (2023) and Runge et al. (2023), the current standard for exploring the interactions between climate variables have been large physics-based simulation models, programmed with equations derived from observational data. However, these models demand high computational resources and are out of reach for many people. They also require specific assumptions to be made, which may hinder their reliability. Therefore, as suggested in Runge et al. (2023), we will apply purely data-driven causal models to quantify the relationships between climate variables. Such models would serve as a ”double-check” to simulation models, in addition to requiring less computing power.

In our causal analysis, We will focus specifically on human-related activity that are well-known to be causes of climate change. This includes measures of global warming, greenhouse gas emissions, energy production, forestation, crop yields, and transportation usage among other variables. We will use two methods to perform causal inference, Neural Network models and Structural Causal Model Frameworks (SCMs). We chose Recurrent Neural Network models (RNNs) based on their suitability when working with time series data and ability to detect highly nonlinear relationships, which we expect to find in our data. SCMs are a dedicated causal framework that seek to represent variables as nodes in a directed graph. An edge from node x to node y means that x has been determined to cause y. SCMs have scarcely been applied in the field of climate change research, so this project will also serve to analyze their effectiveness within the earth system.

The data we are working with is gathered from various sources online that specialize in earth system studies and climate change research, and are all averaged over the globe. The dataset for global temperature is gathered from the Surface Temperature Analysis

website (GISTEMP-Team (2024)) collected by Goddard Institute for Space Studies (GISS) under the National Aeronautics and Space Administration (NASA). The datasets used for CO2 and Methane (CH4) Emissions are gathered from the Global Monitoring Laboratory website (X. Lan and Thoning (2199), X. Lan and Dlugokencky (2199)) and collected by The Global Monitoring Division of the National Oceanic and Atmospheric Administration's (NOAA) Earth System Research Laboratory. The dataset used for electricity generation was gathered from the Energy Information Administration (EIA) website (EIA). In terms of data cleaning, the only major manipulation is the merging of the datasets with indices of [year,month] and removing data where not all variables are present, or in other words, cutting all datasets to match the dataset which started the latest.

2 Methods

2.1 Time-Delayed Causal Relations with Neural Network Implementations

With the assistance of the python package PyTorch, we trained several recurrent neural networks on a large portion of the dataset, leaving some recent data as a validation set. This involved optimizing both the number of layers in our Recurrent Neural Network and the ideal lag value. This model is used to determine Granger causality between variables. We determine causality by including a "filtering" layer at the beginning that uses L1 regularization to discard variables deemed unnecessary to getting a good prediction; this method of discovering Granger causality was first described in Horvath, Sultan and Ombao (2022). We retrain the same model many times to reduce the influence of randomness, and averaged the resulting filter values for a more stable outcome. Additionally, we will use the same RNN architecture (trained on the full dataset) to get the residuals when predicting for each variable; these residuals are later used in detecting instantaneous causality with SCMs.

2.2 Instantaneous Causal Relations with CD-NOD Algorithm and Structural Causal Models

Using the *causallearn* python package, the entire dataset was run through a CD-NOD algorithm to find instantaneous causal relations between each variable. This can then be visualized through a node graph. Using the causal links found from the CD-NOD algorithm, an additive noise model (ANM) was applied to each causal link to discover the direction of the link between each node. The ANM returns probabilities of causal direction between each given variable, and thus an α of 0.01 will be used to determine enough probable cause for a causal relation. The detection of instantaneous relations is important, since our data has been averaged on a monthly basis. One month is enough time for our climate variables to affect one another, but this change would appear to be instantaneous in our data.

2.3 Combined Time-Lagged and Instantaneous Causal Relations

We are also able to combine RNNs and SCMs to create a model that can detect both time-lagged and instantaneous causal relations. First, we fit a RNN to the full dataset, which will attempt to predict one time step into the future using a specified number of lags. Then, we find the residuals of the RNN's predictions and plug this in as input to a SCM. The SCM, in determining causal relations between residuals, will be able to find the instantaneous causal relations that our time-lagged model misses.

Future extensions of this project may also be able to use our models to predict forward in time. This is an inherent feature of a RNN, and is also possible with SCMs as long as we fit functions that compute a node's output value given all of the node's parents. These can be computed in many ways, such as with a Gaussian process regression, or even smaller neural networks. SCMs also inherently allow us to perform an intervention by fixing a node to be a particular value, and predicting the effect on the other variables. For example, we may predict how global temperature might change if global CO2 emissions stopped increasing.

3 Results

3.1 Time-Delayed Causal Relations with Neural Network Implementations

Figure 1 presents the prediction performance for each feature using our trained neural network model.

Figure 2 presents a heat map of Granger Causality importance values of each feature against one another. This indicates how significant a feature can be in predicting a corresponding feature. The directionality of the relationship reads as row -> column. For example, the importance of petroleum production in predicting global temperature is 0.2.

3.2 Instantaneous Causal Relations with CD-NOD Algorithm and Structural Causal Models

Figure 3 presents a node link diagram displaying the results of causal relations found by the CD-NOD algorithm. In instances where directions are discernible, an arrow indicates causal direction, otherwise the direction of the relationship is not immediately clear.

When an Additive Noise Model is applied to Methane and Petroleum data in order to find the direction of relation, the p-values for both directions return $<.01$

$(Methane -> Petroleum, Petroleum -> Methane) = (0.0, 1.29 * 10^{-13})$

When accounting for the possible confounder of "Time", the new additive noise model produces similar results.

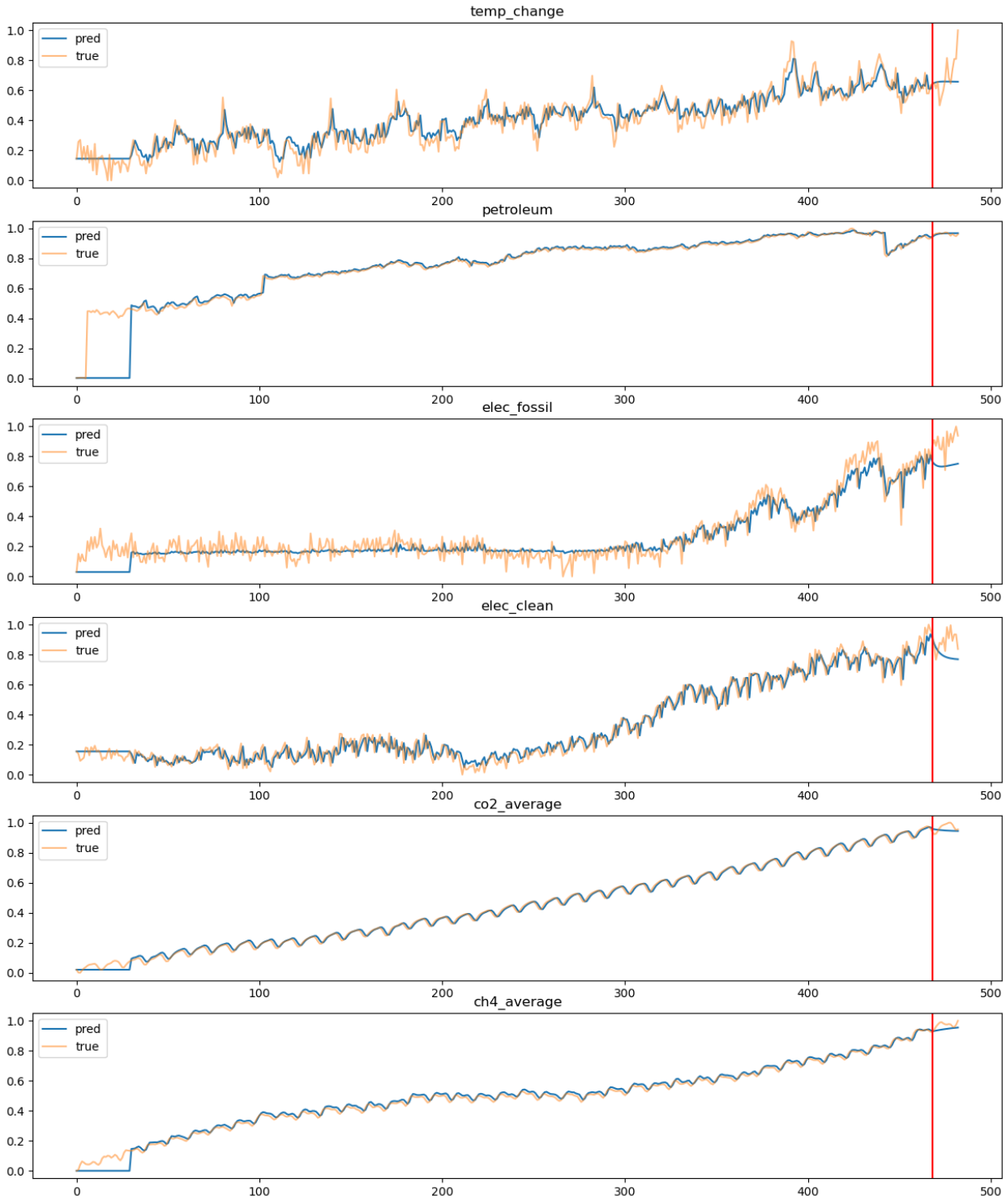

Figure 1: The accuracy of neural network prediction for each variable.

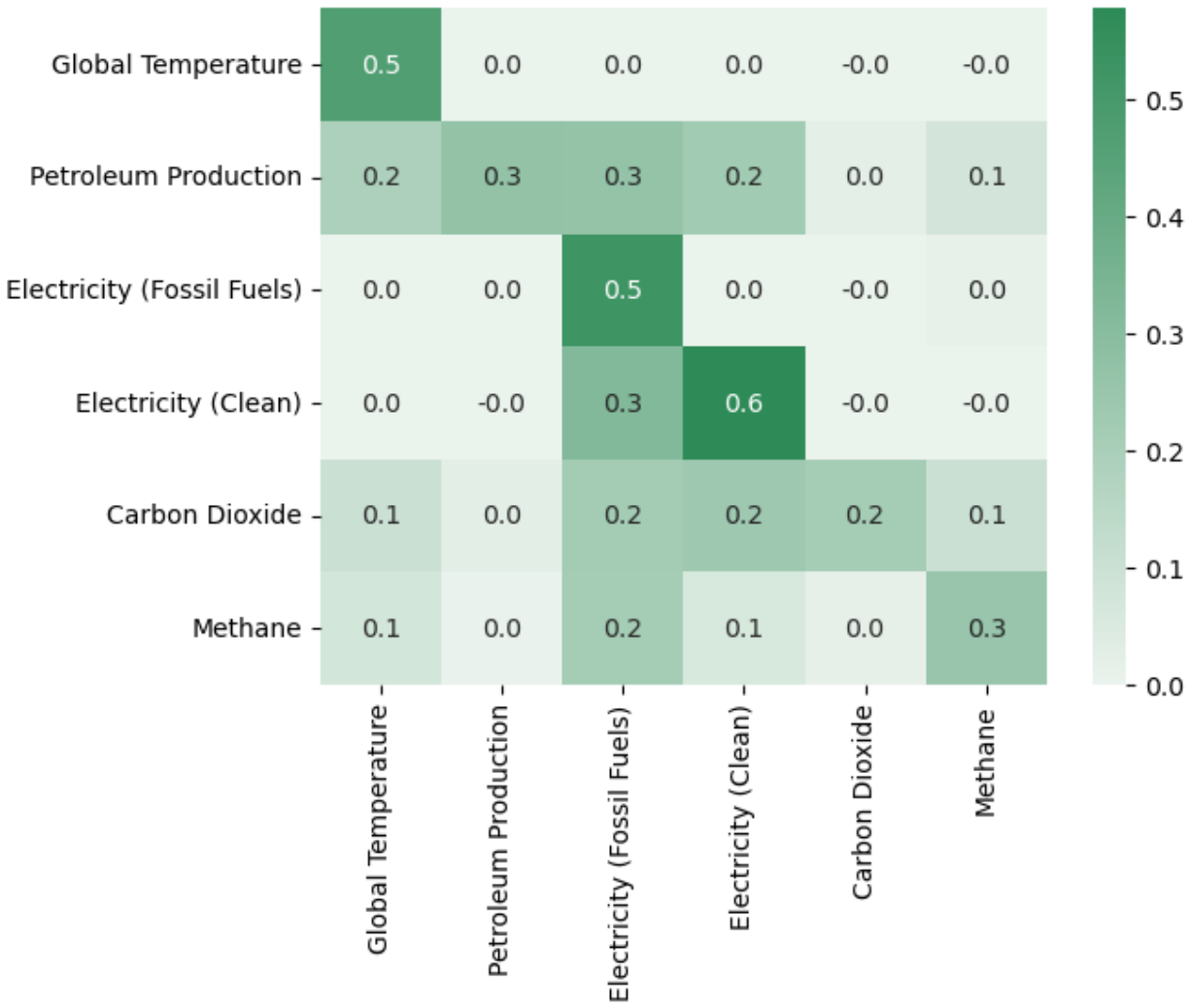

Figure 2: A heat map of each feature's "importance" in affecting another feature.

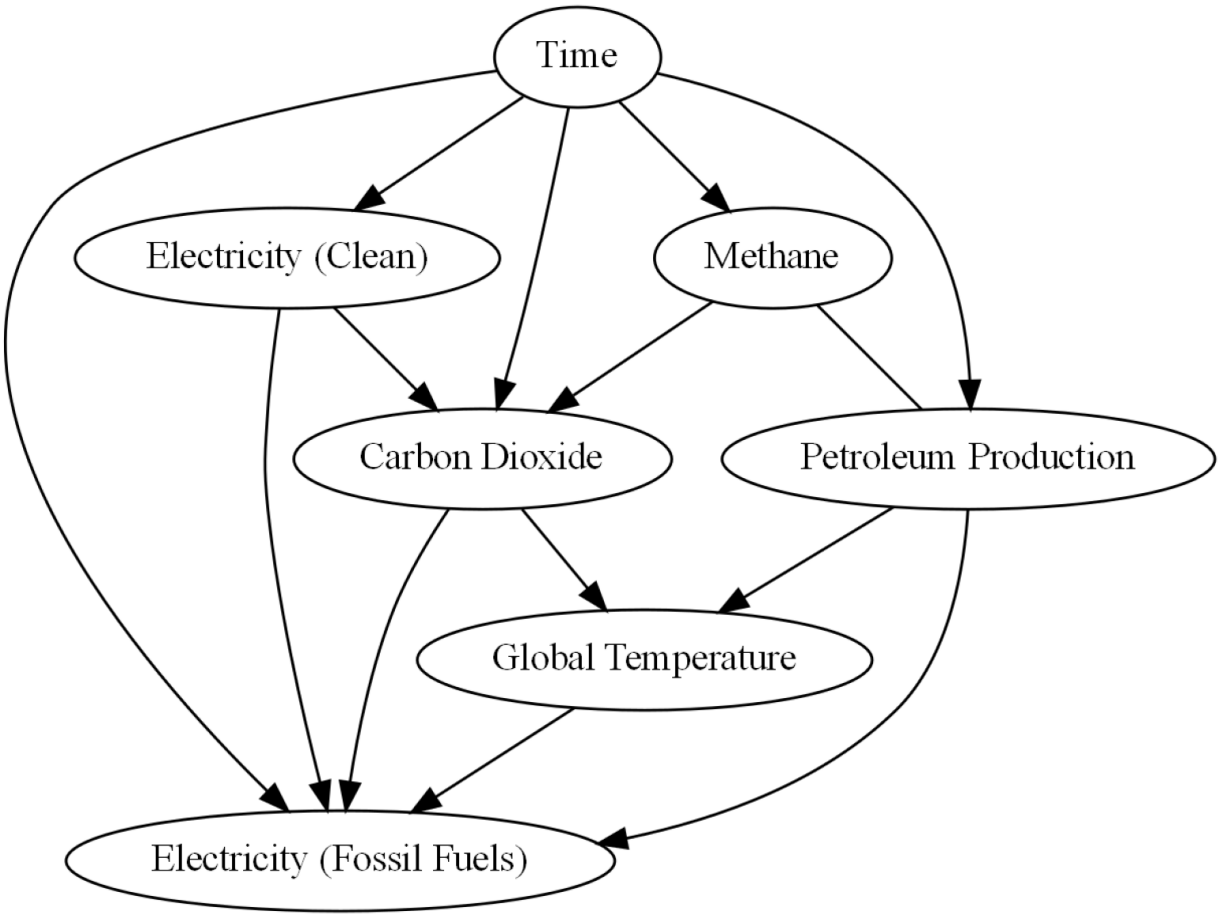

Figure 3: A node-link graph of the causal relations and directions found using CD-NOD algorithm.

$(Methane -> Petroleum, Petroleum -> Methane) = (0.0, 0)$

These results do not make any sense, as the CD-NOD algorithm predicts a link between *petroleum* and *methane* but the ANMs do not; many of the causal relations also do not make sense. However, this behavior is expected. The CD-NOD algorithm does not account for any time lagged causality, and can only consider instantaneous causality.

3.3 Combined Time-Lagged and Instantaneous Causal Relations

Figure 4 presents a node link diagram displaying the results of causal relations found by the CD-NOD algorithm on residuals of the data found from the neural network implementation.

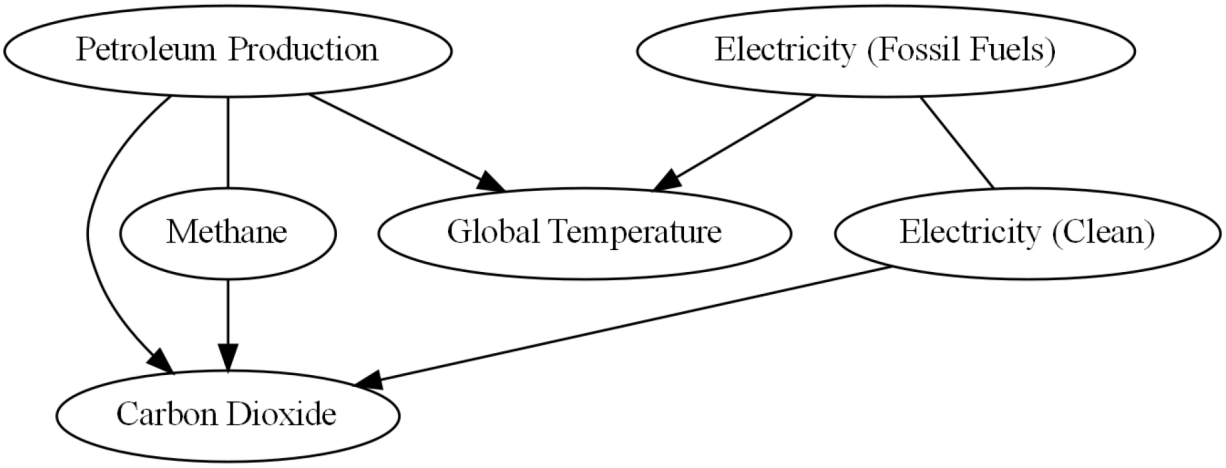

Figure 4: A node-link graph of the causal relations and directions found using CD-NOD algorithm.

When an ANM is applied to the relationship between Electricity (Clean) and Electricity (Fossil Fuels), only the p-value for Clean Electricity causing Fossil Fuels Electricity returns <0.01

$(Clean -> FossilFuels, FossilFuels -> Clean) = (0.0001, 0.0444)$

When an ANM is applied to the relationship between Electricity Petroleum Production and Methane, only Methane causing Petroleum returns a p-value of <0.01.

$(Petroleum -> Methane, Methane -> Petroleum) = (0.0221, 0.0)$

This means that $FossilFuels -> Clean$ and $Petroleum -> Methane, Methane$. These results mostly make sense, as fossil fuel and clean energy production are naturally correlated, and petroleum consumption is indeed a cause of methane emissions.

4 Discussion

4.1 Time-Delayed Causal Relations with Neural Network Implementations

Through testing multiple recurrent neural network structures, we found that less complex models provided the lowest training error since; a high complexity, large neural net would lead to over fitting. The lag value we settled with was 30 as it struck a balance between improving model performance on training data and determining Granger causality, however with larger datasets, the optimal lag value would increase.

As seen in figure 1, the neural network implementation does well in predicting on training data, even capturing certain patterns, but may also be a sign of over-fitting. The portion past the red line indicates validation data which. While the predictions are close to the true values, they are not as specific as the training set predictions.

From the neural network model, we can establish Granger causality for each variable in relation to the others. In figure 2, we are able to visualize the importance of each feature in predicting another feature. From our experiments, we find that a value of $>= 0.01$ implies Granger causality. From the importance heat map, we can draw some interesting conclusions. The values on the diagonal ($feature_n -> feature_n$) are notably higher, which makes sense, as using the feature itself would prove useful in predicting its own future values. We can also see that Petroleum Production, Carbon dioxide, and Methane are important features almost across the board in predictions. On the other hand, Global temperature, Electricity (Fossil Fuels), and Electricity (Clean) seem to be stronger dependent features than predictor features. For many of these, interactions, they make sense with previous knowledge, such as Carbon Dioxide, Methane, and Petroleum Production all being important factors in Global Temperature prediction. Another example is Petroleum Production being a very strong factor in Electricity generated through fossil fuels (such as petroleum). Some entries prove to be slightly confusing however, like Carbon Dioxide and Methane being important factors in predicting Electricity generated from fossil fuels since, through prior knowledge such as EPA (2199), this relationship should be reversed. This is likely a spurious correlation, due to external factors not captured in the dataset.

4.2 Instantaneous Causal Relations with CD-NOD Algorithm and Structural Causal Models

From the CD-NOD algorithm, we are able to discover and visualize causal relationships in our data. Unlike the previous model, which is suited for time-lagged relations, this algorithm detects instantaneous causal relationships. Here we see a much different picture of more closely interconnected data than the previous model. Notably, the inclusion of the Time variable is also able to be included as an intuition of if the flow of time has causal relation to the other features.

Once again, many of these relations do align with our previous knowledge of the earth

system, such as Carbon Dioxide being a factor in Global Temperature. Time being an important factor in all features (aside from global temperature) makes sense as all features are nonstationary and shift over time. Even a confusing relationship from the last model is fixed here, where Petroleum Production directs to Electricity generated by fossil fuels. However, this model does seem to have its own anomalies, such as Global temperature and Clean Electricity generation causing Electricity generated by fossil fuels.

As per our methods, any relationships without clear causal direction can be analyzed with an additive noise model to find the direction. When the relationship between Methane and Petroleum Production is analyzed with an ANM alone, we found that both directions produce a p-value < 0.05, implying no relation. The result is similar when the ANM accounts for the possible confounding variable of Time. These results, although poor, are expected.

4.3 Combined Time-Lagged and Instantaneous Causal Relations

Lastly, the combined Time-Lagged and Instantaneous relation node-link diagram clearly has a much simpler system than the previous two models. It also has by far, the most intuitive conclusions. Petroleum production having relations with CO2, and methane (without clear direction) makes sense along with Global Temperature slightly more indirectly. Electricity generated from fossil fuels also being a factor in Global temperature makes sense too, however, its relation to Clean Electricity is slightly confusing. From the ANM analysis of directionality here, it finds causal direction from Fossil Fuels -> Clean Electricity. This is most likely a result of confounding variables not included in the data, or a high correlation between the two variables. The other relation that requires a directionality test between Petroleum and Methane returns Petroleum -> Methane, which also aligns with background knowledge.

5 Conclusion

In this research, we wished to find causal relations within the Earth system with specific focus on human-caused data. Here, we used three different methods and presented their unique findings. In this, we also explored the use of Structural Causal Models within the Earth system through the use of Additive Noise Models. These proved to be very useful, especially in the combined model. It would be more correct to say that the use of SCM was consistent and helpful, but is very susceptible to confounding variables. With a massive system as big as Earth's, with so many important features and relations, it is notable to state that the models produced here are magnitudes simpler than the real system. Here we are limited to the data that humans are able to gather and only until they have started collecting them. Nonetheless, these models were able to pick up relations that make sense within our understanding and earlier research. Some causal relations are even consistent throughout all three models, such as Petroleum Production being an important factor in Global Temperature. Some Relationships are completely different in each model, such as Electricity generated from fossil fuels having no relation whatsoever (Time-Lagged), to Global Tem-

perature causing Fossil fuel Electricity (Instantaneous), and Fossil Fuel Electricity causing Global Temperature (Combined). Going forward, there are many steps to improve our models. The most important one is gathering more features to use in our dataset to create a more complex system, closer to what the true Earth system is. As time goes on, humans continue to collect data, and find new features to measure and publish. Not only with the continual gathering of data in real time improve any updated version of these models, but adding more features can help explain relationships with confounding variables.

References

(EIA), U.S. Energy Information Administration's., "The Monthly Energy Review." [Link]

EPA., "Sources of Greenhouse Gas Emissions." [Link]

GISTEMP-Team. 2024. "GISS Surface Temperature Analysis (GISTEMP)." *NASA Goddard Institute for Space Studies*. [Link]

Horvath, Samuel, Malik Shahid Sultan, and Hernando Ombao. 2022. "Granger Causality using Neural Networks."

Runge, Jakob, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. 2023. "Causal inference for time series." *Nature Reviews Earth & Environment* 4(7): 487–505. [Link]

Song, Jiecheng, and Merry Ma. 2023. "Climate Change: Linear and Nonlinear Causality Analysis." *Stats* 6(2): 626–642. [Link]

X. Lan, K.W. Thoning, and E.J. Dlugokencky., "Trends in globally-averaged CH4, N2O, and SF6 determined from NOAA Global Monitoring Laboratory measurements." [Link]

X. Lan, P. Tans, and K.W. Thoning., "Trends in globally-averaged CO2 determined from NOAA Global Monitoring Laboratory measurements." [Link]

Appendices

A.1 Data Access

Global Temperature: https://data.giss.nasa.gov/gistemp/ We used the "Global-mean monthly, seasonal, and annual means" dataset. Notably, this dataset updates monthly to add, but also correct previous months, which may cause slight, but ultimately inconsequential variations when downloaded at different times.

CO2 Emissions: https://gml.noaa.gov/ccgg/trends/gl_data.html The dataset is titled "Globally averaged marine surface monthly mean data"

CH4 Emissions: https://gml.noaa.gov/ccgg/trends_ch4/ The dataset is titled "Globally averaged marine surface monthly mean data"

Energy Generation (Clean, via Fossil Fuels): https://www.eia.gov/totalenergy/data/monthly/ The dataset titled "Primary energy production by source" under the "Energy overview" section

Petroleum Production: https://www.eia.gov/international/data/world The dataset titled "Monthly petroleum and other liquids production"

A.2 Website

https://homkeen.github.io/DSC180B-Q2-Capstone/

A.3 Github

For direct access to our code: https://github.com/HomKeen/DSC180B-Q2-Capstone