

Final Paper

STOR 320.03 Group 5

December 7, 2025

Introduction

As our world becomes increasingly data-driven, it has never been more critical to examine what climate-related data reveals about our rapidly changing environment. Climate change represents one of the most pressing challenges of the 21st century, with far-reaching implications for ecosystems and economies around the world. While the scientific consensus on climate change caused by humans is well-established, understanding the relationships between key climate variables and using them to predict future trends is still a very active area of research.

This climate science revolves around three fundamental indicators: global land temperatures, sea levels, and atmospheric greenhouse gas concentrations. These variables form an interconnected system where changes in one can drive changes in the others. Greenhouse gases such as carbon dioxide (CO₂), methane (CH₄), and nitrous oxide (N₂O) trap heat in the atmosphere which leads to rising global temperatures. As temperatures increase, thermal expansion of seawater and melting of ice contribute to rising sea levels. However, the strength of these require investigation careful quantitative investigation.

Our research seeks to answer the question: **what correlations can we observe between global land temperatures, sea levels, and greenhouse gas concentrations, and what do projections of these variables look like over time?** This question is important for a lot of people. For policymakers, understanding which greenhouse gases have the strongest correlation with temperature increases can inform emission reduction strategies. For coastal communities, knowing the relationship between past temperatures and future sea level rise helps guide infrastructure decisions. For the everyone else, clear evidence of these relationships can provide a data-driven foundation for climate awareness.

We approach this question through statistical modeling and time-series analysis of historical climate data spanning from 1992 to 2013. By examining correlation patterns, constructing predictive models, and investigating potential time lags between variables, our goal was to reveal real, actionable insights. For instance, does sea level respond immediately to temperature changes, or is there a delay as oceans absorb heat over time? Among the major greenhouse gases, which exert the strongest influence on global

warming? Can we use these historical relationships to project future trends with reasonable confidence?

We also investigate how the general public sentiment towards climate change has evolved over time. Public discourse on climate change is increasingly shaped by social media, where opinions spread rapidly across geographic and ideological boundaries. Understanding how people discuss climate change online, therefore, is crucial for policymakers and scientists.

The central question for this analysis is: **how do environmental conditions and social demographics relate to public stances on climate change, as expressed on Twitter?** Does the user's geographic location play a meaningful factor in this as well? How can we (as data scientists) best convey this information? To tackle these questions, we combine geolocation data and sentiment scores to investigate how climate related beliefs vary across spatial/social groups.

Data

Our analysis draws upon three primary datasets spanning the period from 1992 to 2013. This 22-year window provides sufficient observations to examine both trends and relationships in the Earth's climate system.

The first dataset contains historical average land temperature data originally compiled by Berkeley Earth and accessed through Kaggle. Each observation includes the year, country, average temperature in degrees Celsius, and an uncertainty measure representing the 95% confidence interval. We computed global annual averages by aggregating temperature measurements across all countries for each year.

The second dataset provides measurements of greenhouse gas concentrations through the Annual Greenhouse Gas Index (AGGI), maintained by the National Oceanic and Atmospheric Administration (NOAA). This dataset tracks concentrations of major heat-trapping gases including carbon dioxide (CO₂), methane (CH₄), nitrous oxide (N₂O), and various fluorinated compounds such as chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), and hydrofluorocarbons (HFCs). The AGGI represents a summary index quantifying the cumulative radiative forcing of all these gases relative to a 1990 baseline.

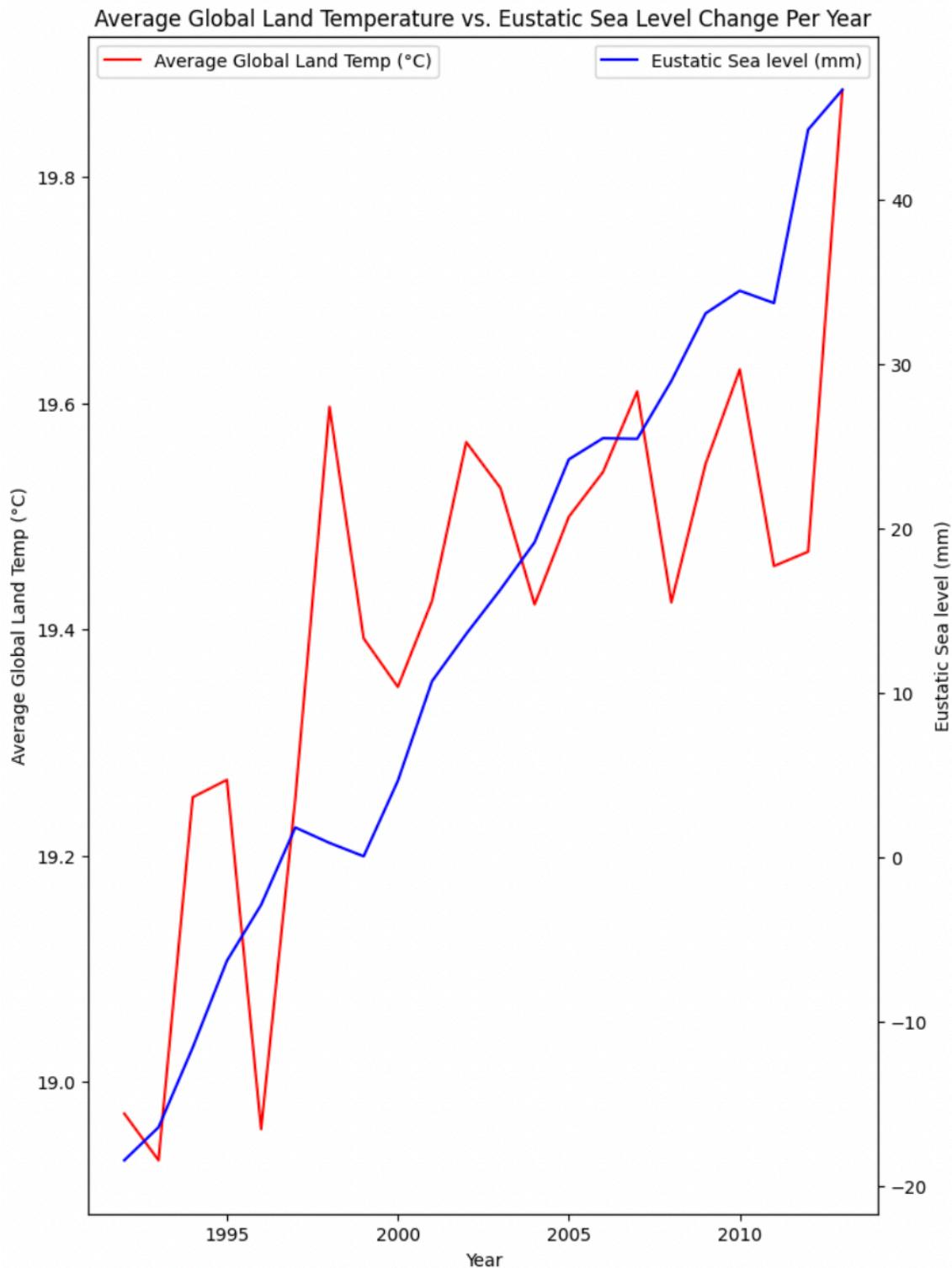
The third dataset describes eustatic sea level—the total volume of all ocean water if contained in a single basin—measured as anomalies compared to the mean sea level in 1993–1999. This data is maintained by researchers at the University of Colorado and reports sea level in millimeters above or below the baseline, with seasonal cycles removed.

Our data preparation involved several key steps. We first checked each dataset for missing values and removed incomplete observations to maintain data integrity. The land temperature data was converted from year-month-day format to simple years, while the sea level data's decimal year format was converted to standard datetime objects for consistent analysis. We then merged the three datasets using an inner join on the year variable, retaining only years present in all three sources. This resulted in our final 22-year dataset from 1992 to 2013. We used backward-filling to handle a few remaining gaps where earlier values were missing, particularly in the smoothed temperature and sea level variables.

We created several derived variables to facilitate analysis. Because the fluorinated compounds (CFCs, HCFCs, HFCs) represent a related family of gases regulated under the Montreal Protocol, we combined them into a single "Other Gases" variable. We also applied min-max normalization to create scaled versions of each variable, transforming them to a 0-to-1 scale to enable fair visual comparisons across variables with different units and magnitudes.

	df_merged														
	Year	CO2	CH4	N2O	CFC*	HCFCs	HFCs*	Total	CO2_eq_ppm	AGGI	AGGI_Change	AverageTemperature: Celsius	gmsl_mm	land_smooth	sea_level_smooth
0	1992	1.364	0.495	0.137	0.343	0.022	0.003	2.365	433.0	1.028	1.2	18.972007	-18.465000	19.075991	-11.150180
1	1993	1.375	0.495	0.138	0.346	0.024	0.004	2.382	434.0	1.035	0.7	18.930800	-16.455441	19.075991	-11.150180
2	1994	1.399	0.497	0.140	0.348	0.025	0.004	2.413	437.0	1.049	1.4	19.251702	-11.599324	19.075991	-11.150180
3	1995	1.427	0.500	0.141	0.349	0.027	0.005	2.449	440.0	1.064	1.5	19.267070	-6.317871	19.075991	-11.150180
4	1996	1.455	0.501	0.144	0.350	0.028	0.005	2.483	443.0	1.079	1.5	18.958375	-2.913265	19.075991	-11.150180
5	1997	1.472	0.502	0.146	0.350	0.030	0.006	2.506	445.0	1.089	1.0	19.250351	1.799818	19.131659	-7.097216
6	1998	1.512	0.506	0.149	0.350	0.031	0.007	2.556	449.0	1.111	2.1	19.596783	0.859818	19.264856	-3.634165
7	1999	1.545	0.509	0.152	0.350	0.033	0.008	2.596	452.0	1.128	1.8	19.392100	0.045455	19.292936	-1.305209
8	2000	1.563	0.509	0.156	0.349	0.035	0.008	2.620	454.0	1.139	1.0	19.349202	4.642382	19.309362	0.886842
9	2001	1.587	0.509	0.158	0.348	0.036	0.010	2.647	456.0	1.151	1.2	19.425303	10.710306	19.402748	3.611556
10	2002	1.617	0.509	0.160	0.347	0.038	0.011	2.682	459.0	1.166	1.5	19.565523	13.587389	19.465782	5.969070

This table shows a sample of our merged dataset, illustrating how the key climate variables align by year. All variables show upward trends over the study period, with greenhouse gases, temperature, and sea level all increasing from 1992 to 2013.



This figure displays the parallel trends in global land temperature and sea level rise over our study period. Both variables show clear upward trajectories, with temperature increasing from approximately 18.97°C to 19.88°C , while sea level rises from -18.47mm to 46.72mm relative to the 1992 baseline. The visual synchronicity provides initial evidence of correlation, which we explore through statistical modeling in the Results section.

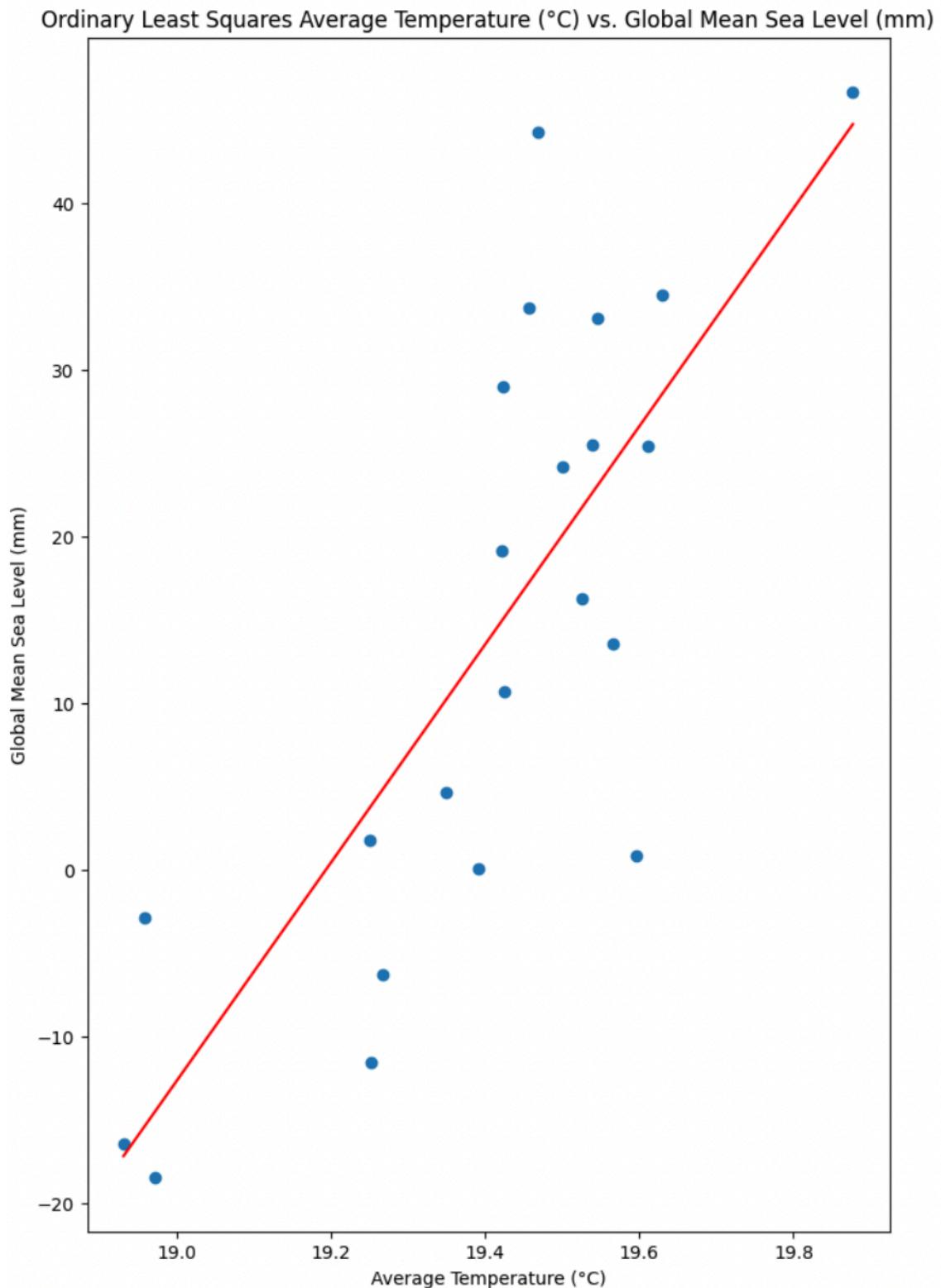
To address the public sentiment question, we utilized a dataset consisting of geolocated tweets related to climate change, sourced from a publicly available Kaggle repository. Each observation includes the tweet timestamp, geographic coordinates, topical classification, sentiment score, stance toward man-made climate change (believer, denier, neutral), user-inferred gender, aggressiveness label, and local temperature deviation from historical climate averages. After filtering for valid latitude and longitude values, the final working dataframe contains 5307538 entries (individual tweets) and occupies roughly 182 MB in memory after optimization.

Results

To answer our research question about correlations between global land temperatures, sea levels, and greenhouse gas concentrations, we employed multiple statistical approaches including correlation analysis, linear regression, and vector autoregression modeling.

Correlation Between Temperature and Sea Level

We began with a simple linear regression to quantify the relationship between global land temperature and sea level. This model treats sea level as the dependent variable and average temperature as the independent variable.



The regression analysis reveals a strong positive relationship with an R-squared value of 0.626, indicating that approximately 62.6% of variance in global sea level is explained by land temperature. The coefficient of 65.49 means that for every 1°C increase in global land temperature, sea level rises by approximately 65.5 millimeters on average. With a p-value near zero, this relationship is highly statistically significant.

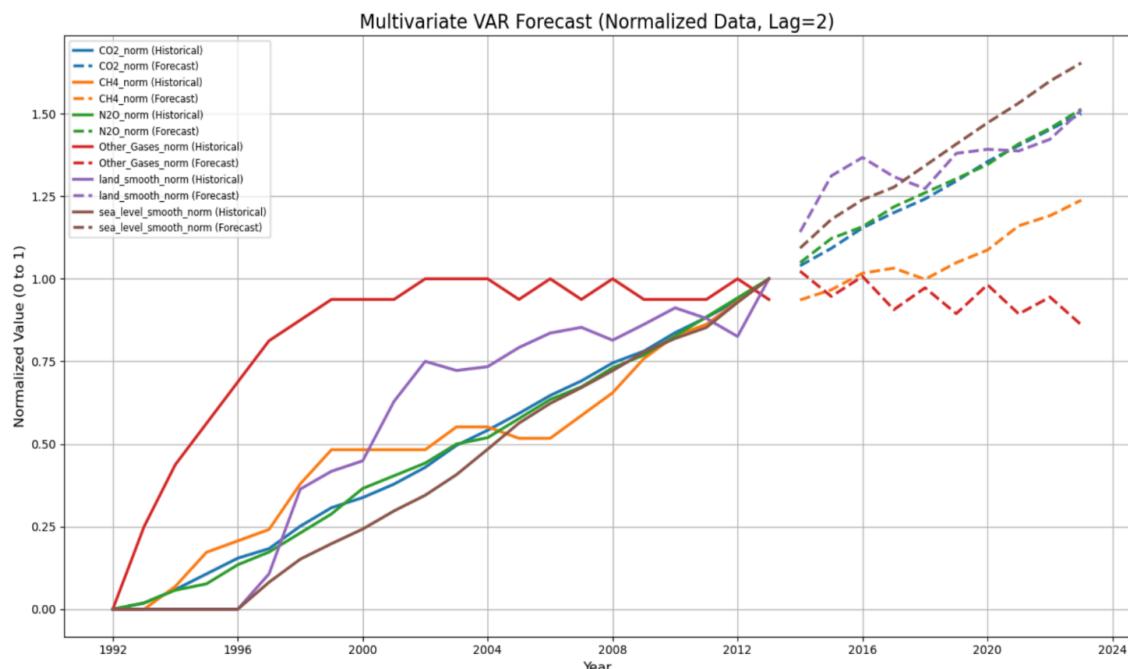
Contemporaneous Correlations

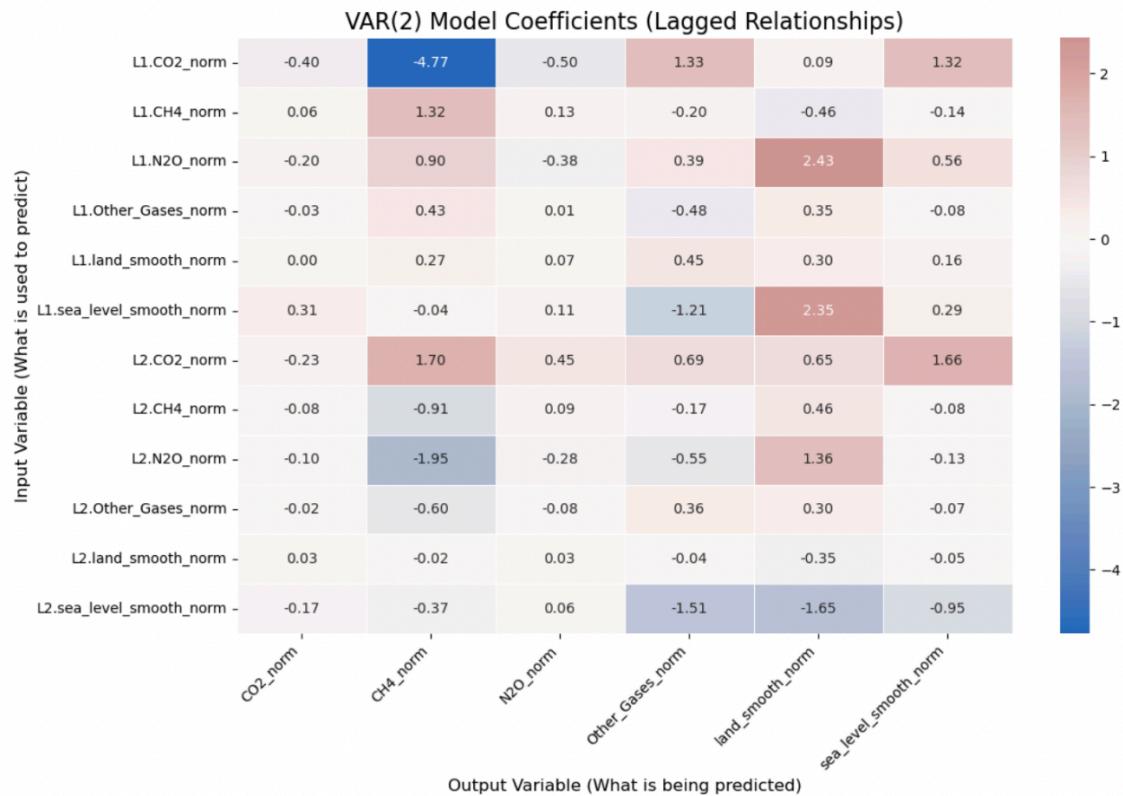
To examine how variables relate within the same year, we created a correlation heatmap based on annual changes rather than raw values. This approach avoids spurious correlation from shared upward trends.

Analysis of the year-to-year changes in climate variables revealed two key findings regarding their relationships. First, the Annual Greenhouse Gas Index (AGGI) showed only a weak immediate correlation (0.28) with land temperature and sea level metrics, suggesting that climate variables do not respond instantaneously but often involve time delays. Second, the fluorinated gases (Other_Gases_norm) exhibited negative correlations with other variables. This reflects the successful impact of the Montreal Protocol in phasing out these compounds, even while other major greenhouse gases continued their upward trend.

Vector Autoregression Forecast and Lag Analysis

To capture time-lagged relationships and project future trends, we employed a Vector Autoregression (VAR) model. VAR treats multiple variables as an interconnected system where each variable's future depends on past values of itself and other variables. Our model uses two years of historical data (lag 1 and lag 2) to forecast the period from 2014-2023.





The forecast shows distinct patterns: Other_Gases continues declining due to regulation, land temperature increases dramatically to 2016 then moderates, major greenhouse gases (CO2, CH4, N2O) all increase consistently, and sea level rises steadily throughout the projection period.

The VAR coefficient matrix reveals which past values most strongly predict future temperature and sea level. For land temperature, the most significant predictors are N2O from one and two years prior, sea level from one year prior, CO2 from two years prior, and CH4 from two years prior. For sea level, the dominant predictors are CO2 from both one and two years prior, and N2O from one year prior.

This lag structure reveals important climate dynamics: sea level rise is most directly influenced by recent CO2 emissions (1-2 year delay), while land temperature responds to a more complex mix of gases with N2O playing an unexpectedly prominent role. The different response timescales suggest that atmospheric temperatures can shift relatively quickly, while ocean mass and volume changes accumulate more gradually but track closely with recent radiative forcing.

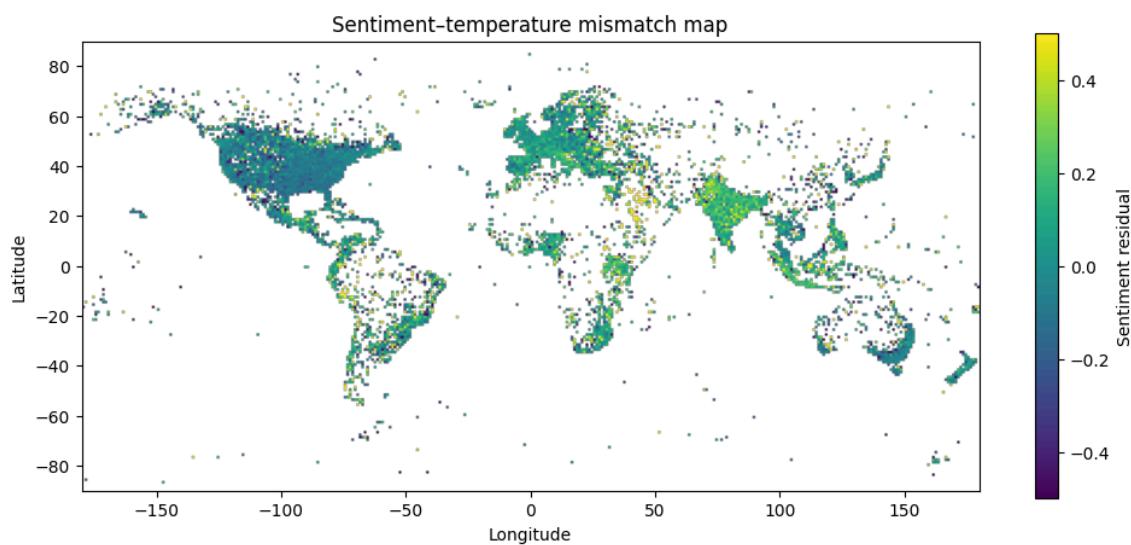
In summary, our analysis demonstrates clear statistical relationships between greenhouse gases, temperatures, and sea levels. The linear regression confirms a strong temperature-sea level connection, while the VAR model exposes the time-lagged nature of climate dynamics. Sea level responds primarily to CO2 with a 1-2 year delay, while land temperatures are influenced by multiple gases—particularly N2O and CO2—with

varying lag periods. These findings provide quantitative evidence for the interconnected, time-dependent nature of the climate system.

Sentiment Visualization

The supplied density map (on Kaggle) shows that climate related tweets are heavily concentrated in North America, Western Europe, and parts of South and Southeast Asia, with especially strong activity in the eastern United States and Western Europe. Large portions of Africa, South America, and Central Asia show much lower tweet density, reflecting both population differences and unequal access to Twitter.

Overall, the pattern highlights a strong geographic imbalance in where online climate discourse is most active. This spatial clustering is important to keep in mind when interpreting global trends from the data.

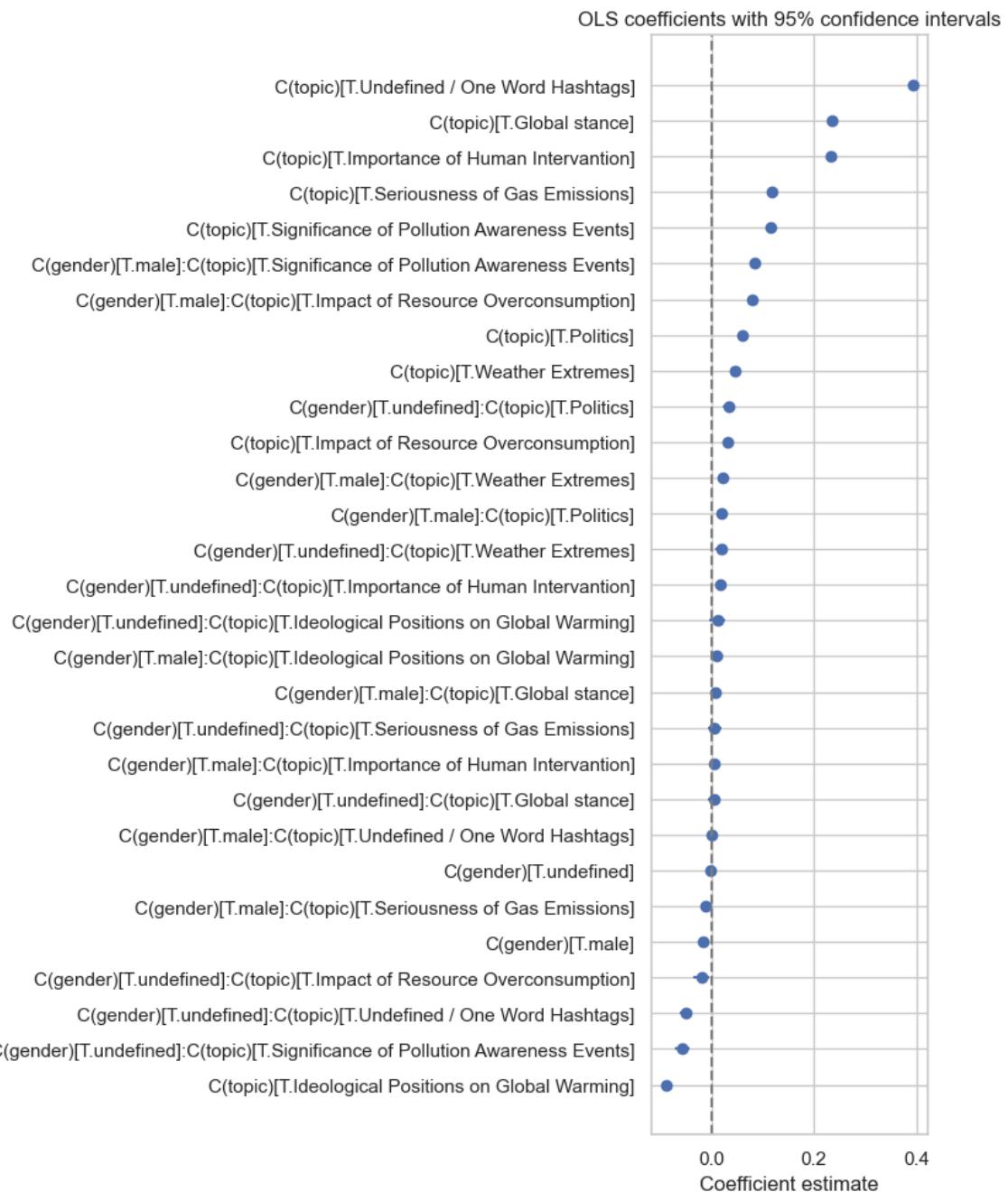


This map shows where public sentiment diverges from what local temperature anomalies alone would predict. Bright positive areas indicate regions where people express more positive sentiment than expected given local climate deviations, while darker areas indicate the opposite. There is a clear cultural bias within this effect, as sentiment tends to be bucketed by country. We note that the majority of the United States expresses a strongly negative residual, which cannot be explained just from temperature events alone. The Global South is more varied: areas like the Middle East and South Asia express positive deviations whereas Southeast Asia expresses more negative residuals.

Sentiment Inference

To understand how user predictors (like location or gender) relate to their sentiment, I fit an OLS regression with sentiment as the dependent variable and categorical predictors for stance, topic, and gender, plus a continuous temperature term. The model explains about 8.3% of the variation in sentiment, which is small but meaningful given the

noisiness of Twitter text. The results show a very clear ideological divide: deniers score 0.18 points lower in sentiment than believers, indicating substantially more negative emotional tone. Neutral users score slightly higher than believers. Topic categories also show strong differences. For instance, tweets tagged as "Undefined / One Word Hashtags" have the highest positive sentiment (+0.36), while ideological or conflict-heavy topics like "Ideological Positions on Global Warming" trend negative. Gender has no statistically meaningful effect, as both male and undefined-gender coefficients are small and non-significant. Temperature shows a small but statistically significant positive relationship with sentiment. The regression suggests that **ideological position and topical framing matter far more for sentiment than gender or environmental conditions.**



We next examined how stance varies geographically by binning tweets into global latitude bands. The chi square test is extremely large ($\chi^2 = 44154, p < 0.001$), indicating a very strong association between latitude and stance. Believers consistently dominate across nearly all latitude bands, but the proportions shift in subtle ways. In mid-latitude regions where most tweets originate (26 to 56 degrees N), believer proportions are in the 73-75% range, with deniers making up 6-8%. Near the equator and southern hemisphere, believer rates remain high but neutral tweets become relatively more common—for example, in the -17 deg to -2.5 deg band, neutrality reaches 43%, the highest anywhere. At the poles, believer rates exceed 80%, though sample sizes are small. These results indicate that climate change belief is globally dominant. **However, the intensity and certainty of belief vary by region, with equatorial areas expressing more neutrality and mid-latitudes showing greater polarization.**

To explore demographic patterns, we analyzed the relationship between gender and stance. The chi-square test again shows a strong association ($\chi^2 = 18821, p < 0.001$). Female users show the highest believer proportion (77.5%), compared to males (73.0%) and undefined users (71.6%). Denial is most common among males (7.8%) and least common among females (5.1%). Meanwhile, neutrality is more common among undefined gender accounts (22.6%). Column proportions reveal that males dominate overall tweet volume in every stance category, which reflects broader platform demographics. Overall, these results show that **women express the strongest alignment with climate change belief while men show more polarization and a higher rate of denial.** Undefined gender users lean more ambivalently.

Finally, to understand how gender and topic jointly influence sentiment, I ran a two-way ANOVA. Topic has by far the strongest effect on sentiment ($F = 44206, p < 0.001$), confirming the regression findings that topical framing is a major determinant of emotional tone. Gender alone has a small but statistically significant effect ($F = 5.0, p = 0.006$), though the effect size is extremely small relative to topic. The interaction between gender and topic is also significant ($F = 173, p < 0.001$), indicating that the emotional tone of climate tweets differs by gender depending on the specific topic. This means, for example, that male and female users may sound similarly positive on one topic but diverge strongly on another. Overall, the ANOVA shows that **sentiment is shaped primarily by what people are talking about, secondarily by who is speaking, and meaningfully by the interaction between the two.**

Males are disproportionately represented in the following topics: Donald Trump versus science, politics, and weather extremes. Females, meanwhile, are relatively more active in topics related to human intervention, pollution awareness events, and resource overconsumption. Undefined gender accounts show higher proportions in low volume topics.

Conclusion

Our investigation into the relationships between global land temperatures, sea levels, and greenhouse gas concentrations revealed several key findings. Linear regression analysis demonstrated a strong correlation between temperature and sea level, with approximately 62.6% of sea level variance explained by temperature alone. For every 1°C increase in global land temperature, sea level rises by approximately 65.5 millimeters on average—a concrete measure of how atmospheric warming translates into ocean expansion.

The correlation heatmap further illustrated that greenhouse gases and climate impacts have weak contemporaneous relationships, indicating that climate variables respond with delays rather than instantaneously. Our Vector Autoregression model captured these dynamics by modeling time-lagged effects across multiple variables. The VAR forecasts project continued increases in CO₂, CH₄, and N₂O through 2023, along with steady sea level rise and ongoing warming. The coefficient analysis showed that sea level responds most strongly to CO₂ emissions from one to two years prior, while land temperature is influenced by multiple gases—particularly N₂O and CO₂—with varying delay periods.

These findings have practical implications. For policymakers, our results highlighting N₂O's underappreciated role alongside CO₂'s dominant influence can inform emission strategies to reduce greenhouse gas influence on temperature. For coastal planners, the 1-2 year lag between emissions and sea level response informs them that we are already committed to some future rise in sea level based on past emissions.

Furthermore, we find that climate discourse on Twitter is not evenly distributed across space, topics, or social groups. Instead, it reflects strong geographic clustering, ideological polarization, and measurable relationships between environmental conditions and public emotional response. The presence of temperature anomalies alongside shifts in sentiment suggests that real-world climate conditions may directly influence online belief expression.

Across all sentiment analyses, the data consistently show that beliefs about climate change on Twitter are strongly shaped by stance, topic, and geography, while demographic factors like gender play a smaller role. Sentiment is overwhelmingly more positive among believers, and deniers reliably express more negative sentiment regardless of topic. Geographic patterns further reinforce this divide, with belief-heavy regions showing the strongest climate-supportive sentiment and denier activity concentrated in narrow latitude bands. Topic patterns reveal that discussions about global stance, human intervention, and pollution awareness drive the most positive sentiment, whereas ideological debates trend negative. Finally, the interaction effects

between gender and topic confirm that the subjects people engage with matter far more than who they are.

For policymakers, these findings have important implications for climate communication and misinformation intervention strategies. Understanding where denial is concentrated, which topics provoke the most aggressive rhetoric, and how environmental conditions shape sentiment can help tailor region-specific outreach efforts and improve the effectiveness of science communication campaigns. More generally, this study demonstrates how large scale social media data can be integrated with environmental variables to study the social dimensions of climate change in real time.

Our analysis also has important limitations. Our dataset spans only 22 years, which is relatively short for climate analysis. Climate systems operate on century-long timescales, and patterns observed over two decades may not capture longer-term dynamics. With only 22 observations, the VAR model is at risk of overfitting. Additionally, our models assume linear relationships and may miss nonlinear feedback mechanisms that become important at higher temperature thresholds.

Future research could extend this work by incorporating additional decades of data, exploring nonlinear modeling techniques to capture threshold effects, and validating forecasts against observed values from 2014-2025. Ultimately, our findings contribute quantitative evidence linking greenhouse gas emissions to measurable climate changes, with the time-lagged relationships reminding us that today's emissions commit us to tomorrow's impacts.