Exploring the relationships between CO2, temperature and sea level in relation to climate change.

José Luis Ferres.

Biologist, Data Scientist,

- Github: JFerrs

- LinkedIn: https://www.linkedin.com/in/jferrs/

Abstract

This report details an analysis and modelling of Global Mean Sea Level (GMSL) rise between 1993 and 2023, examining its relationship with annual per capita CO2 emissions and global temperature anomalies. The main objective was to understand the dynamics among these climatic variables and to develop robust predictive models. The methodology included comprehensive exploratory data analysis (EDA), the implementation of linear regression models (both direct and through a chaining approach), and a Prophet time series model with exogenous regressors. Results revealed an extremely high correlation between temperature anomalies and GMSL. The linear regression model using temperature and year as predictors demonstrated high accuracy (RMSE of 2.37 mm, R^2 of 0.91), slightly outperforming the Prophet model (RMSE of 2.74 mm, R^2 of 0.88), though both models showed excellent fit. Multicollinearity between CO2 and the time factor was identified and discussed within the linear models. 10-year projections, performed with both approaches, indicate a continuous and concerning rise in Global Mean Sea Level, underscoring the necessity for climate action. This analysis provides a foundation for future research into the impact and prediction of climate change.

1. Introduction

Despite all the scientific evidence, many deny climate change, questioning human influence on it. It's true that climate change has always occurred, but studies show that human activities, especially those related to industries, accelerate this process due to the release of greenhouse gases.

However, the scale and speed of the changes observed in recent decades are unprecedented in recent geological history, distinguishing current warming from past natural cycles. The accumulation of gases like carbon dioxide in the atmosphere, stemming from the burning of fossil fuels, has intensified Earth's natural greenhouse effect, trapping more heat and leading to a sustained increase in global temperatures.

One of the most tangible and concerning consequences of this warming is the rise in Global Mean Sea Level (GMSL). This phenomenon, driven by the thermal expansion of water as oceans warm and by the melting of glaciers and polar ice sheets, poses a direct threat to coastal communities, marine ecosystems, and global infrastructure. Understanding the relationship between CO2 emissions, global temperature, and GMSL is fundamental for effective policy formulation and adaptation to a changing climate future.

The present study seeks to contribute to this vital understanding, using historical data to empirically analyse the interconnections among per capita CO2 emissions, global temperature anomalies, and GMSL. Through the application of various modelling techniques, including linear regression and advanced time series analysis with Prophet, this project aims not only to identify causal relationships but also to offer data-driven projections for sea level in the coming decades, providing a clearer insight into the impact of human activity on our planet.

2. Data sources and preprocessing

Data for this project were sourced from reputable scientific and public databases:

- Global Temperature Anomalies (Annual_Anomaly_Lowess_C). Sourced from the NASA Goddard Institute for Space Studies (GISS).
- Annual CO2 Emissions per capita. Obtained from Our World in Data.

- Global Mean Sea Level (GMSL mean). Acquired from the NASA Sea Level Change team.

Robust data analysis hinges upon meticulous preprocessing, a critical phase that ensures dataset integrity and readiness. This typically commences with a comprehensive assessment to understand data structure and identify issues. Key steps include the judicious management of missing values (e.g., NaN imputation or removal), the pruning of extraneous columns to optimize dimensionality, and the standardization of data types to ensure consistency. Finally, integrating disparate datasets through merging on common keys unifies the information, yielding a refined, reliable, and analysis-ready foundation for subsequent rigorous exploration.

3. Exploratory Data Analysis.

3.1 Visual analysis.

From the first DF graph in figure 1 is obtained. A clear increase in anomalous temperatures is observed, starting with negative values at the end of the 19th century, and becoming positive since the middle of the 20th century. Also, grafic shows a warming from 1940 onwards with a linear trend and larger error bars in recent decades, which indicates increasingly extreme anomalies.

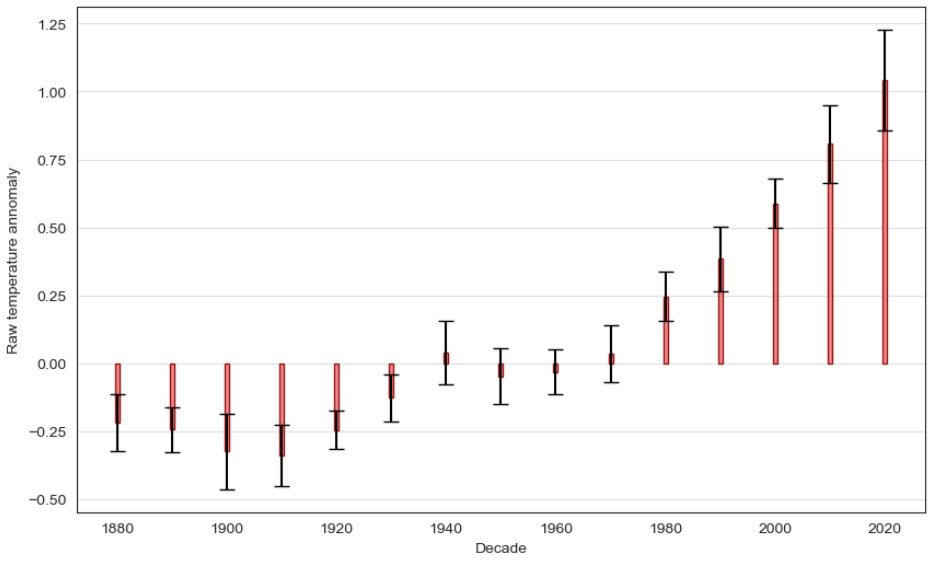

Fig. 1. Temperature anomalies by decades from 1880

Second DF shows CO2 emission data. In this case, the global entity from among all those in the table has been used in order to obtain an overview of the data over time. CO2 values are very low from the beginning of the records until 1850, when they begin to rise almost exponentially. This increase is due to the advent of the industrial revolution, which began in the mid-18th century, increasing the levels of carbon dioxide emitted.

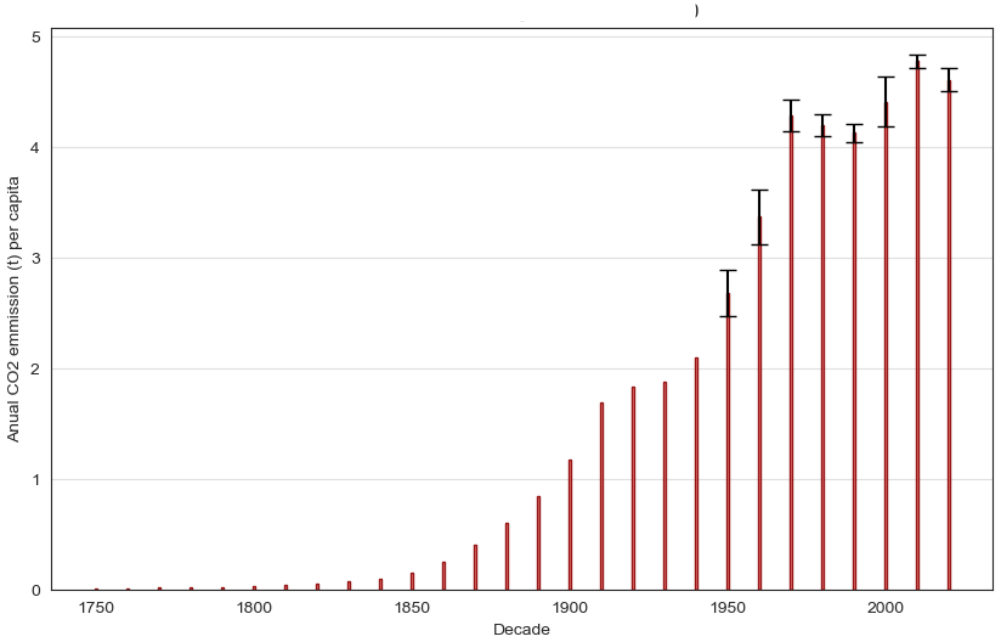

Fig. 2. CO2 emissions per capita by decades.

The last DF reveals data on sea level variation over the last 30 decades. It starts with negative values, but this does not indicate a sea level below zero; rather, these data represent deviations from a pre-established reference average. This means, the zero value in this time series corresponds to an average sea level calculated during a specific (more recent) base period, causing previous levels (along with their anomalies) to appear negative. Similarly, a clearly linear upward trend can be observed in each decade.

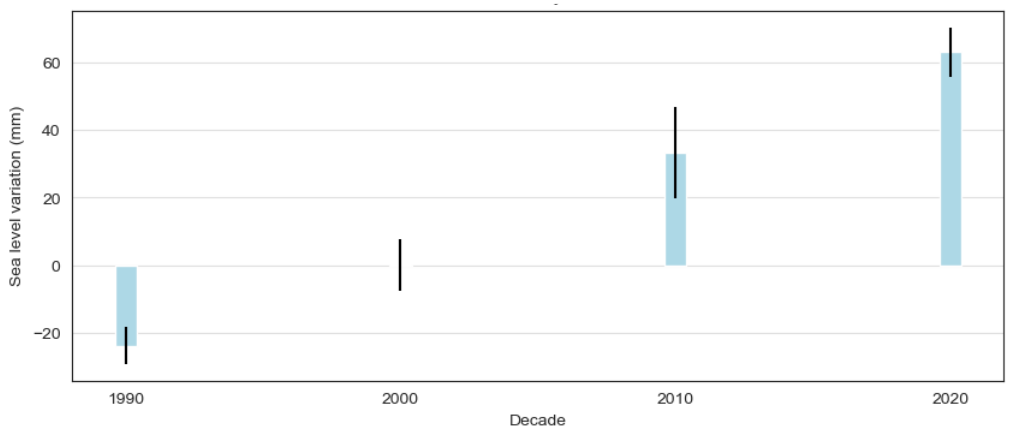

Fig. 3. Sea level variation in mm by decades

Initially, all three DF seems to have an initial correlation due to the shape of the graphs and how the values grow over time. To verify the connection between the different data sets, we will conduct statistical tests that support our initial hypothesis of correlation.

3.2 Statistical analysis

To understand how different parts of the climate system relate to each other, Pearson correlation analysis serves as a key statistical tool. This method is ideal for measuring the strength and direction of a linear relationship between two variables. In the study of climate data, such as CO2 emissions, temperature variations, and changes in sea level, where these variables are often observed to follow joint patterns over time, Pearson correlation allows us to confirm if a change in one variable is consistently related to a change in another. A result close to +1 or -1 will indicate a very strong relationship (direct or inverse), helping to strengthen hypotheses about which factors are driving climate change and to quantify the connection between them.

Following the merge of the three DataFrames, a Pearson correlation analysis was conducted. Recognizing that performing this analysis by decade could potentially diminish its statistical power, the correlation was subsequently re-evaluated for each individual year, starting from 1993, the earliest available date for the seaa level variation DataFrame.

The matrix obtained (fig. 4), shows a high correlation between variables, especially for GMSL mean and annual temperature anomalies.

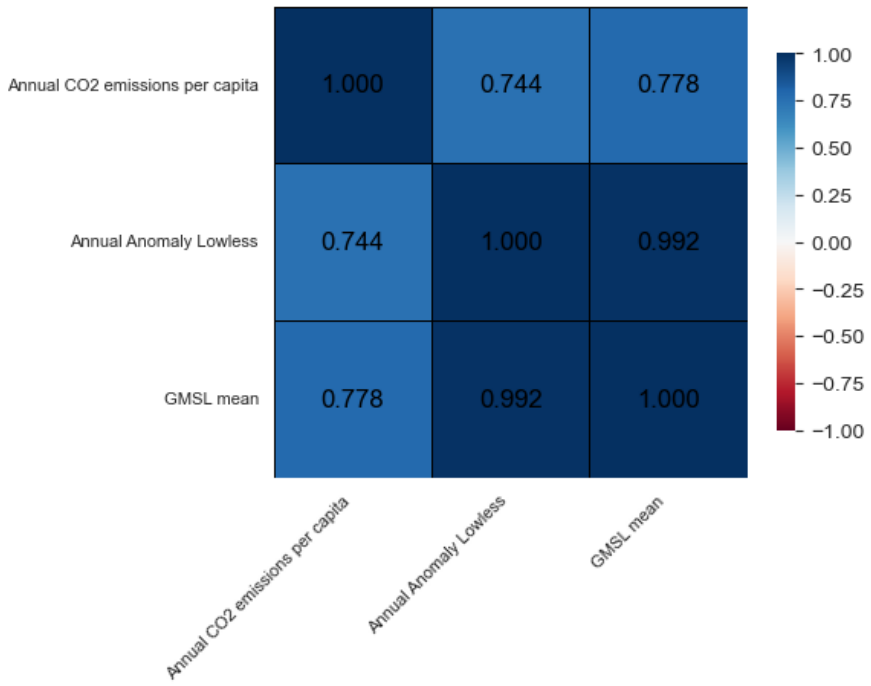

Fig. 4. Pearson correlation matrix of climate variables.

To ascertain the statistical robustness of the observed linear relationships, a Pearson correlation significance test was subsequently conducted for each pair of climate variables. This test is crucial for determining whether the strong correlations found within our sample data are genuinely indicative of a relationship in the broader climate system, rather than merely occurring by random chance.

The results, demonstrating extremely low p-values (< 0.001) coupled with high correlation coefficients, provide compelling statistical evidence. Such low p-values confirm that the probability of observing these strong linear associations by random chance is negligible, thereby affirming the statistical significance and inherent reality of these correlations within the analysed period. This robust statistical support strongly reinforces the scientific understanding of the profound and consistent interconnections between rising CO2 emissions, increasing global temperatures, and escalating sea levels.

Knowing that these different variables are actually correlated, it is now possible to try to create a machine learning model with which to predict the behaviour of some of them.

4. Machine learning models and predictions.

4.1 Linear regression models

To further explore and model the underlying relationships within the climate data, a machine learning approach was adopted, commencing with linear regression as a foundational model. This choice is predicated on its inherent simplicity, high interpretability, and computational efficiency, making it an excellent baseline, particularly given the clear linear trends observed in the preliminary correlation analyses between variables such as year and temperature anomaly.

The efficacy of this linear regression model is quantified by two key metrics: the coefficient of determination (R2) and the Root Mean Squared Error (RMSE). A high R2 value (e.g., nearing 1.0) indicates that a substantial proportion of the variance in the dependent variable (e.g., temperature anomaly) is successfully explained by the independent variable(s) (e.g., year), suggesting a strong fit. Concurrently, a low RMSE value, relative to the scale of the dependent variable, signifies that the model's predictions are, on average, very close to the actual values, implying high predictive accuracy and minimal residual error. Based on these favourable metrics, the linear regression model demonstrates considerable effectiveness in capturing and predicting the observed climatic trends.

The first predictive model (fig. 5), which sought to explain temperature anomalies based on annual CO2 emissions per capita, demonstrated robust performance. The resulting high R2 value of 0.67 indicates that nearly 67% of the variance in global temperature anomalies can be accounted for by the variations in per capita CO2 emissions, suggesting a very strong explanatory power of the model. Furthermore, the Root Mean Squared Error (RMSE) of 0.046, described as low relative to the overall scale of the temperature anomaly data, confirms the model's precision, signifying that its predictions closely align with the observed temperature anomalies. Collectively, these metrics underscore the model's significant ability to

capture and effectively predict the discernible trend of temperature changes as influenced by CO2 emissions.

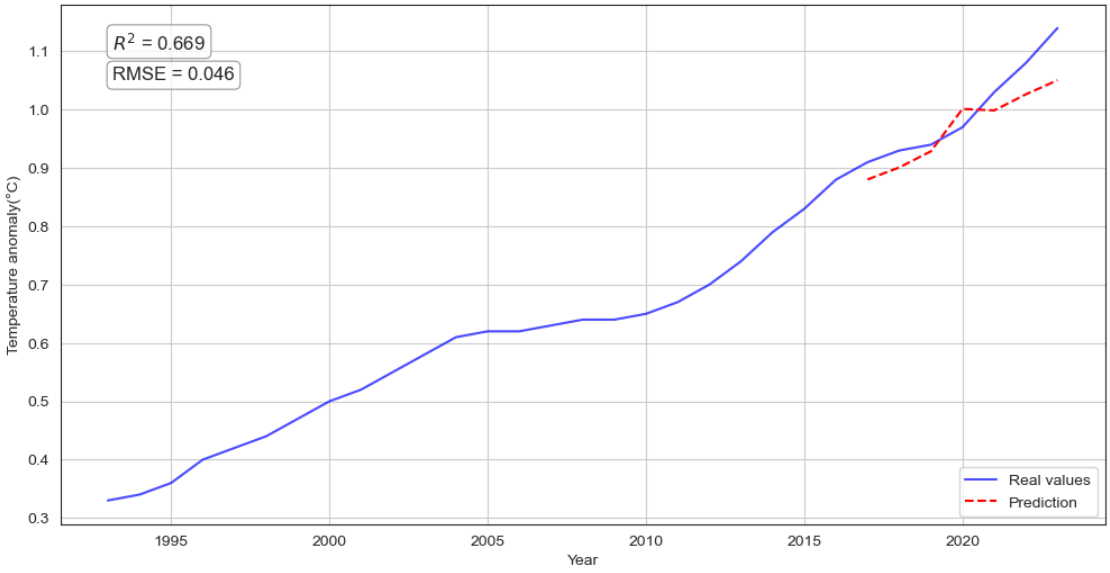

Fig. 5. Temperature anomaly prediction with lineal regression model in base of CO2 emissions per capita

In the second model (fig. 6) the objective is to explain the sea level variation through temperature anomaly values. For this case, a value of R2 is obtained, emphasising the importance of the correlation. Also, a RMSE of 2.3, indicates a low error value, compared to the scale. Both data show a strong linear regression model.

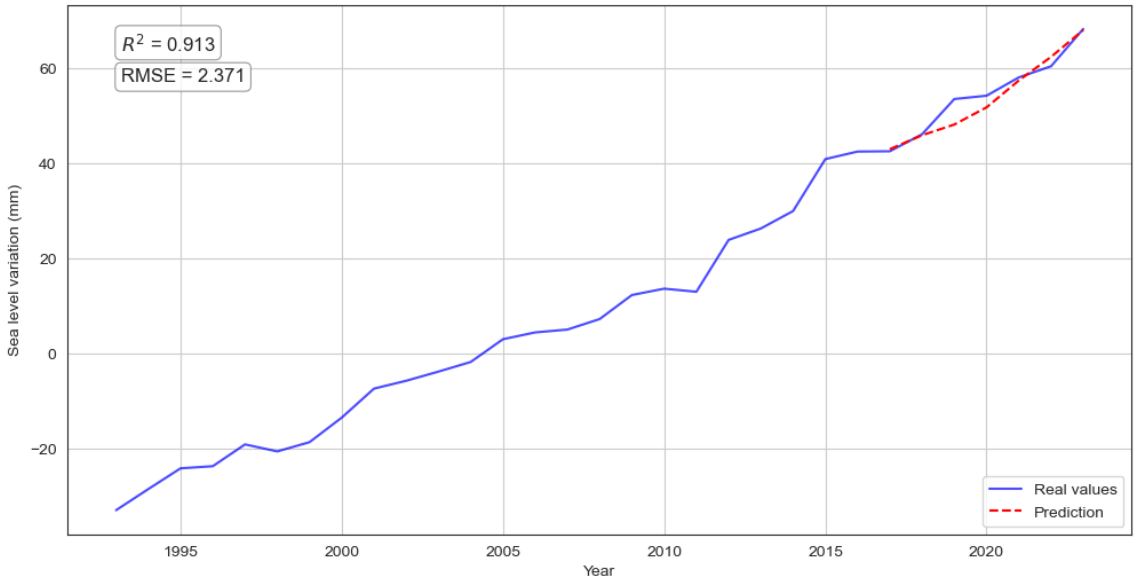

Fig. 6. Sea level variation prediction with linear regression model in base of temperature anomalies.

Last of the linear regression analyses tries to predict sea level variation using CO2 emissions for each year. In this model values of R2 are also high and RMSE error low as in the second model.

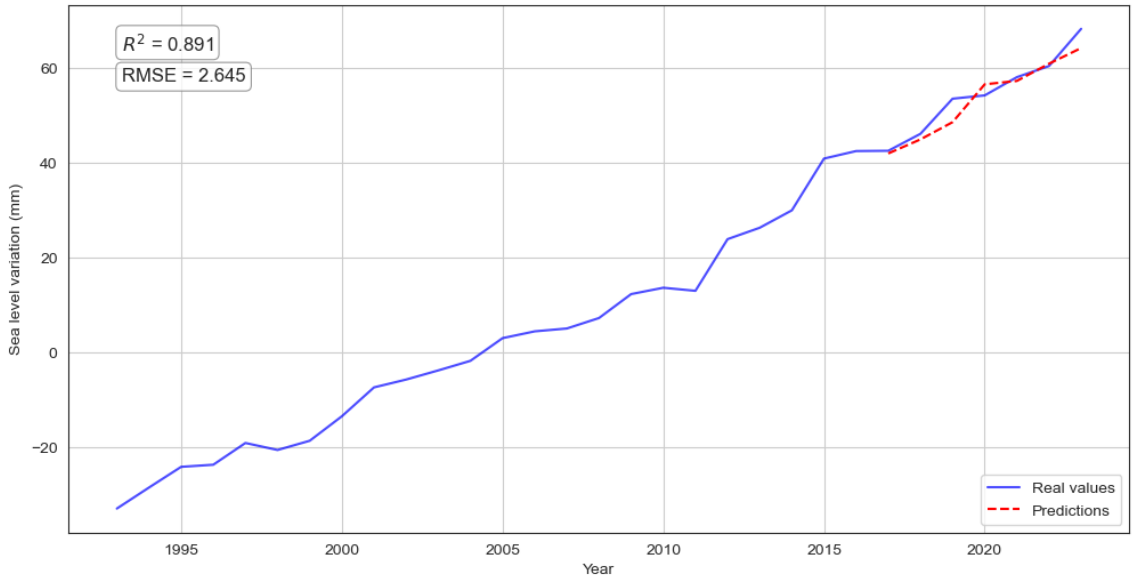

Fig. 7. Sea level variation prediction with linear regression model in base of annual CO2 emissions per capita.

4.2 Prophet model

When analysing environmental time series data, a basic linear regression model offers simplicity but falls short in capturing crucial patterns. Environmental data frequently exhibits pronounced seasonal variations (e.g., daily temperature cycles, annual pollution patterns), evolving long-term trends (due to climate change or policy shifts), and impacts from specific events like extreme weather. To accurately model these complexities and achieve more robust forecasts, I will then use Prophet. This specialized tool inherently handles multiple seasonalities, automatically detects trend changepoints, and allows for the inclusion of important environmental events, making it superior for environmental data analysis.

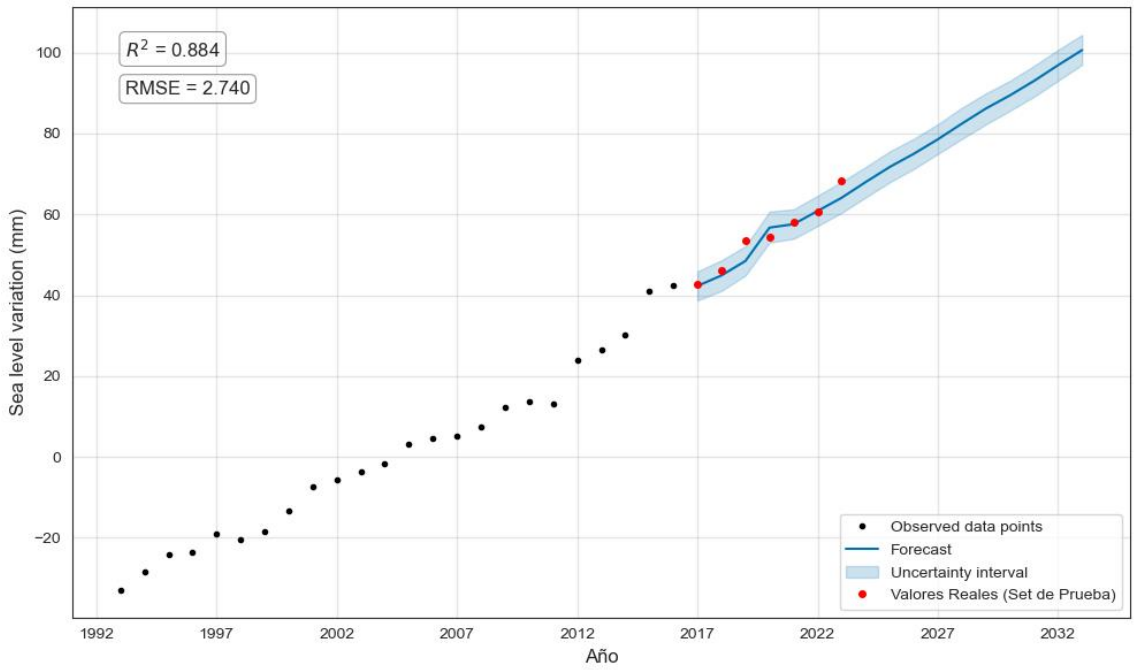

Fig. 8. Sea level variation future predictions using Prophet in base of temperature anomalies.

Prophet model obtains a very similar values or R2 and RMSE than in linear regression, which again supports our initial hypothesis. Also, future values prediction shows a growing linear tendency in case of sea level variation, reaching 100 mm in year 2032.

5. Conclusion

This analysis powerfully reinforces the profound and consistent interconnections between rising CO2 emissions, increasing global temperatures, and escalating sea levels. Our empirical findings demonstrate an extremely high correlation between global temperature anomalies and Global Mean Sea Level (GMSL), with robust statistical significance. This confirms that human activities, particularly industrial emissions of greenhouse gases, are indeed accelerating climate change, driving unprecedented changes in recent geological history.

The predictive models developped—both the chained linear regression and the Prophet time series model—consistently project a continuous and concerning rise in GMSL over the next decade. Reaching predictions of up to 100 mm by 2032, these data-driven projections provide clear, tangible evidence of the trajectory our planet is on. While both models demonstrated excellent fit to historical data, the linear regression model using temperature and year slightly outperformed Prophet in accuracy for the test set (RMSE of 2.37 mm, R^2 of 0.91 vs. RMSE of 2.74 mm, R^2 of 0.88).

The undeniable evidence presented underscores a critical message: the direct threat posed by rising sea levels to coastal communities, marine ecosystems, and global infrastructure is escalating. This study's robust statistical support and data-driven projections serve as a clear call to action. Understanding these relationships is not merely academic; it is fundamental for effective policy formulation and urgent adaptation strategies in the face of a rapidly changing climate future.