Sequential machine learning models to keep climate change under control

Machine Learning course project

Lorenzo Levantesi

A.Y. 2021/2022

Contents

1 Introduction

1.1 What is climate change?

Among the recent worrying bad news that we are absorbing in these recent years, there is one that seems to be not fully understood by most of the people and on which maybe we should care more about, this is the **climate change**.

Climate change explanation can start with the definition of CO_2, i.e., carbon dioxide. CO_2 is a gas present in the atmosphere of the earth, which plays an important role in the **carbon cycle**. This cycle is a delicate dynamic in the earth, given that it defines the carbon dioxide exchanged between the spheres of the earth, which have an almost perfect balance.

This balance has been broken since the second industrial revolution, where the CO_2 produced by the humans has started to grow with a fast rate. This results in an high quantity of CO_2 in earth atmosphere, and, given that the carbon dioxide is a *greenhouse gas*, this means that the energy of the earth's surface is mostly absorbed by the CO_2, resulting in a warmer land and atmosphere.

Some of the most alarming consequences of the increase of the earth's temperatures are:

- **Sea level rise**: As the temperature of sea is rising, land ice melts and increase the expansion of seawater. The sea level rise can cause an increase of flooding, erosion of beaches, and loss of many marshes and wetlands.

- **Increase of hurricanes strength and intensity**: warmer temperatures and rise of the sea level could intensify tropical storm wind speeds, potentially delivering more damage if they make landfall.

- **More heat waves**: All the seasons will tend to get hotter, resulting in heat waves more intense and cold waves less intense and less frequent

- **Increase of wildfires**: Rising of the temperatures increase vegetation flammability.

It seems clear that the rising of the temperatures is not something to take lightly.

1.2 Objective of the project

Given the increasing of availability of data and the high computational power of modern computers, machine learning is becoming really useful in critical scenarios as climate change.

The **objective** of this project is to train a *sequence model* to predict the future CO_2 emissions of the earth. Once obtained this model, another *sequence model* is trained to predict the *anomaly temperatures* of the earth's land and ocean from the CO_2 emissions. Then, a final model is trained to predict the sea level by the anomalies of the land and ocean temperatures.

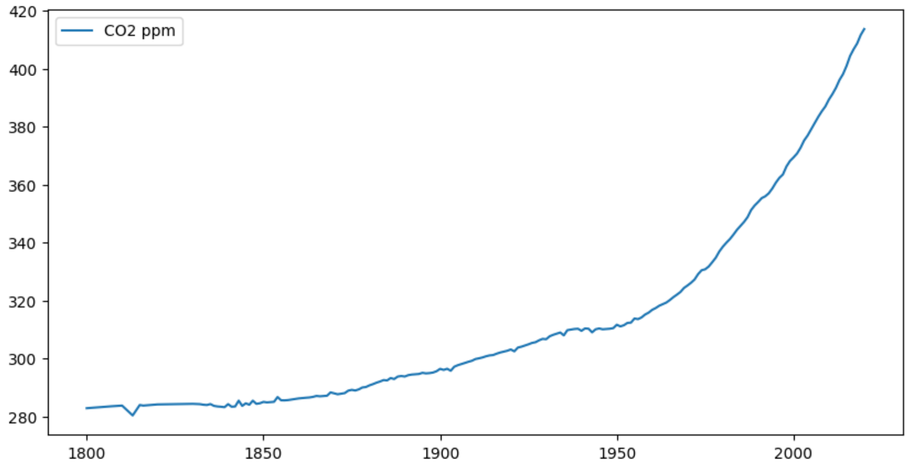

Figure 1: Dataset's CO_2 ppm distribution

These models could be useful to predict future scenarios of the climate change, and then can help to plan a strategy to take action against this problem.

2 Data exploration

All the datasets used in the project are on GitHub.

2.1 CO2 dataset

The first part of the data, for the CO_2 emissions of the earth, has been taken from Ethereidge et al. [1], which contains the yearly historical CO_2 ppm (i.e., parts per million) data between the 1000 and 1958 retrieved from the study of the Law Dome, which is a large ice dome in Antarctica. The most recent measurements of the atmospheric CO_2 concentrations have been taken from the dataset of Dr. Tans and Dr. Keeling, which contains the yearly ppm levels of CO_2, from the 1958 to 2022, measured at the Mauna Loa Observatory in the Hawaii's.

The Figure 1 shows the CO_2 ppm distribution among the years.

Only the data starting from the 1800 are considered, this because we want to let learn the models the correlation between the human activity and the the CO_2 produced, then we consider the data only from the start of the second industrial revolution (i.e., circa 1870), where human activity started to impact significantly the carbon dioxide production.

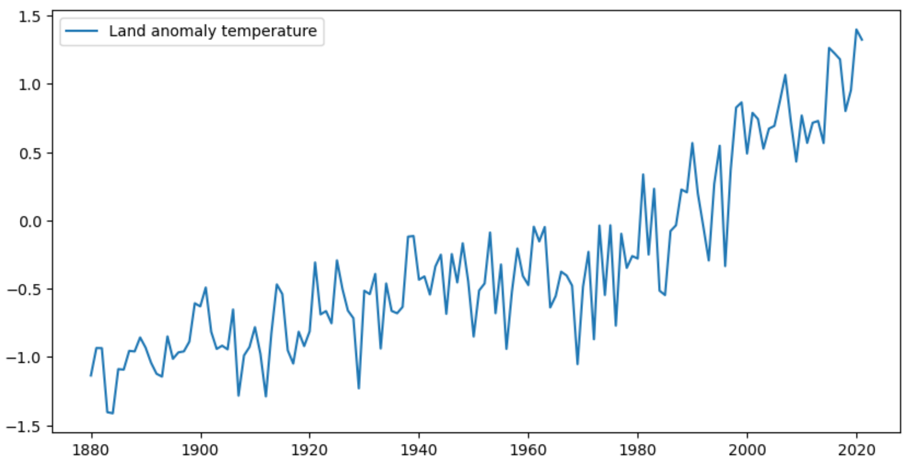

Figure 2: Dataset's land anomaly temperature distribution

2.2 Land and ocean anomaly temperatures dataset

The data for the land and ocean anomaly temperatures are taken from the NOASA Merged Land Ocean Global Surface Temperature Analysis (NOAA-GlobalTemp) dataset.

The dataset contains the anomalies of the temperatures of land and ocean from 1880 to 2021 with respect to the climatology from 1971 to 2000.

The Figure 2 and Figure 3 show the distribution of the two data. Even in this case we can clearly see a rising pattern in the anomaly of the temperatures.

Initially, the dataset contained the monthly measurements of the anomalies. Given that the final models will predict the target values with an annual frequency, for each year the average among the months is computed.

2.3 Sea level dataset

The sea level dataset has been taken from Kaggle, which contains the sea levels from 1880 to 2014. Even in this case the data are on a monthly basis, so, for each year, the average is taken.

The unit measure of the sea level is the *Global Mean Sea Level* (GMSL).

The Figure 4 shows the GMSL values from 1880 to 2014, also in this case there is an increasing trend.

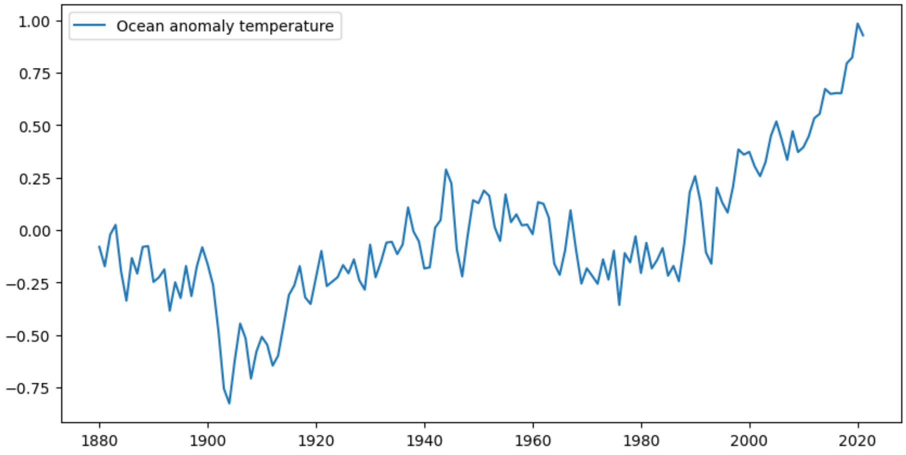

Figure 3: Dataset's ocean anomaly temperature distribution

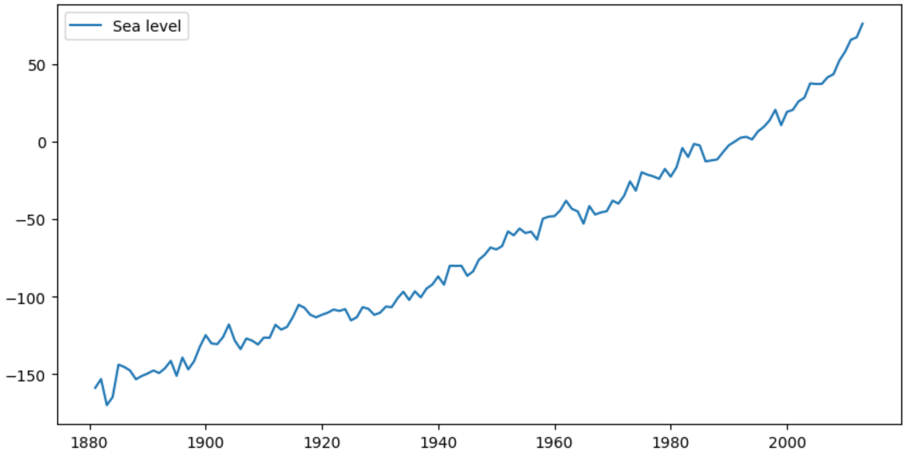

Figure 4: Dataset's GMSL sea level data

3 Model selection

3.1 CO2 prediction

The first task is to train a model that is capable to predict, from a window of 10 years of CO_2 emissions, the CO_2 emission for the next year.

Given the sequential nature of the problem, the models chosen for this task are the **Recurrent Neural Networks** and the **Long Short-Term Memory** models, where the one that performs better will be chosen.

Both the models are *sequential models* and then have a similar way of performing computation. The main difference between them is that the LSTM includes a *memory cell* that lets the model to learn longer-term dependencies. Another important aspect about the LSTM is that they deal with vanishing and exploding gradient problem by introducing new gates, such as input and forget dates, which allow better control over the gradient flow.

This task is started by performing hyperparameters tuning with the *k Fold Cross-Validation* technique on the RNN and LSTM model, where $k = 5$ folds have been used.

The metric used for the evaluation of the models is the *Mean Squared Error* (MSE), which is defined as:

$$\text{MSE } = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the true value and $\hat{y}_i$ is the model's prediction.

The hyperparameters on which the **RNN** has been tested are:

- **Hidden size**: $2, 5, 10$.
- **Number of layers of the model**: $1, 2$.
- **Learning rate**: $0.001, 0.01, 0.1$.

The best hyperparameters found are:

- **Hidden size**: 10.
- **Number of layers of the model**: 2.
- **Learning rate**: 0.01.

which has obtained a validation MSE loss of 0.65.

For the **LSTM**, the hyperparameters on which it has been tuned are:

- **Hidden size**: $30, 70, 100$.
- **Learning rate**: $0.001, 0.01$.

And the best hyperparameters found are:

- **Hidden size**: 30.

- **Learning rate**: 0.001.

which has obtained a validation MSE loss of 1.12.

Even if the LSTMs are in general more powerful with respect to the RNNs, in this task the RNNs performs slightly better. This can derive from the fact that the LSTMs needs a lot of data to learn, where here we have almost 200 temporal windows.

3.2 Land and ocean tempearture anomaly prediction

This second task consists in predicting the anomaly (with respect to the climatology from 1971 to 2000) of the land and ocean temperature for the next year, by having the data from the past 10 years of the CO_2 ppm values.

Also for this task the models chosen are the RNN and LSTM, on which hyperparameters tuning with *k Fold Cross-Validation* is performed, where $k = 5$ folds have been used.

The metric used for the evaluation of the models is still the *Mean Squared Error* (MSE).

The hyperparameters on which the **RNN** has been tested are:

- **Hidden size**: $2, 5, 10$.
- **Number of layers of the model**: $1, 2$.
- **Learning rate**: $0.001, 0.01, 0.1$.

The best hyperparameters found are:

- **Hidden size**: 10.
- **Number of layers of the model**: 2.
- **Learning rate**: 0.1.

which has obtained a validation MSE loss of 0.24.

For the **LSTM**, the hyperparameters on which it has been tuned are:

- **Hidden size**: $30, 40$.
- **Learning rate**: $0.001, 0.01, 0.1$.

And the best hyperparameters found are:

- **Hidden size**: 40.
- **Learning rate**: 0.001.

which has obtained a validation MSE loss of 0.166.

In this task the LSTM have performed slightly better than the RNN. One important thing to notice is that this results may be misleading. In fact the target values mostly range between -1 and 1, so, even if the MSE is low the model can still be not accurate as expected.

3.3 Sea level prediction

This second task consists in predicting the sea level of the next year, by having access to the CO_2 emissions and sea level of the past 5 years. This task is harder then the previous, as the sea level can be influenced by many factors.

Also for this task the models chosen are the RNN and LSTM, on which hyperparameters tuning with *k Fold Cross-Validation* is performed, where $k = 5$ folds have been used.

The metric used for the evaluation of the models is still the *Mean Squared Error* (MSE).

The hyperparameters on which the **RNN** has been tested are:

- **Hidden size**: $2, 5, 10, 20$.
- **Number of layers of the model**: $1, 2$.
- **Learning rate**: $0.001, 0.01, 0.1$.

The best hyperparameters found are:

- **Hidden size**: 10.
- **Number of layers of the model**: 2.
- **Learning rate**: 0.1.

which has obtained a validation MSE loss of 28.07.

For the **LSTM**, the hyperparameters on which it has been tuned are:

- **Hidden size**: $10, 20, 40, 70$.
- **Learning rate**: $0.001, 0.01, 0.1$.

And the best hyperparameters found are:

- **Hidden size**: 40.
- **Learning rate**: 0.001.

which has obtained a validation MSE loss of 4.72.

Also in this case the LSTM have performed better.

4 Evaluation

4.1 CO2 prediction

For the evaluation phase the original dataset has been splitted with the 80% of the data used for training and the other 20% used for test. The RNN has been trained until convergence of the MSE loss, Figure 5 shows the loss distribution during the training phase.

The test loss achieved by the final model is 1.08, which is a pretty good result, as it can be also seen from Figure 6.

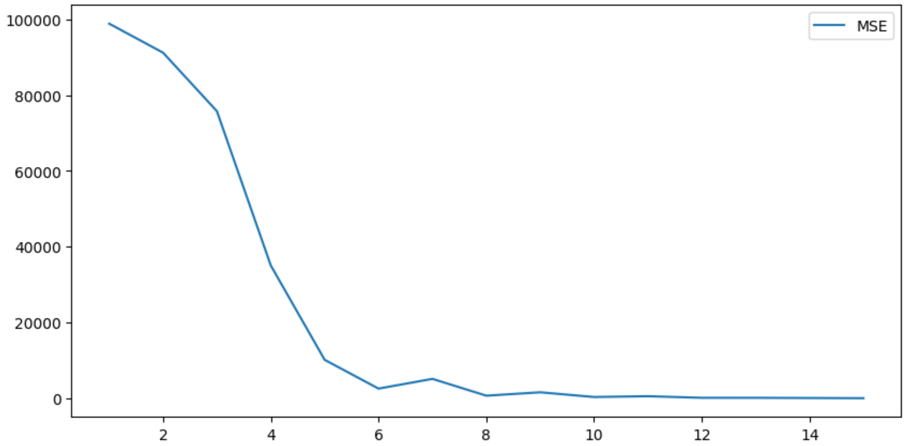

Figure 5: Loss of the training of the RNN

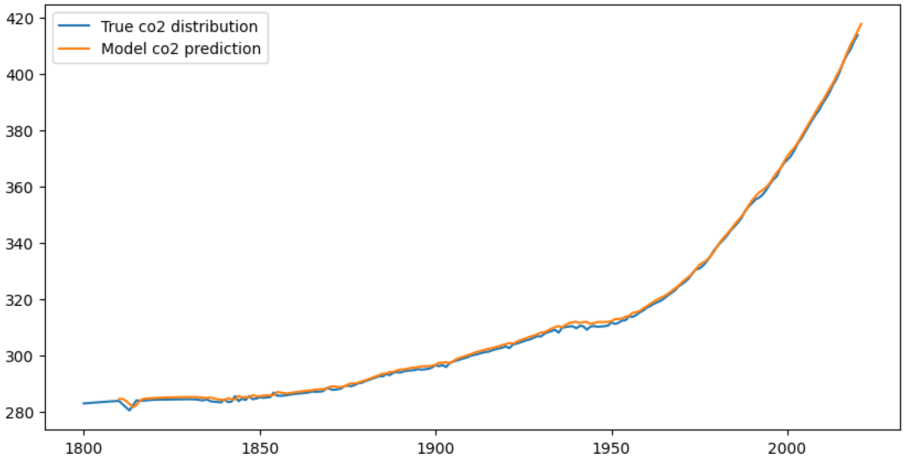

Figure 6: RNN model predictions on future CO2 values

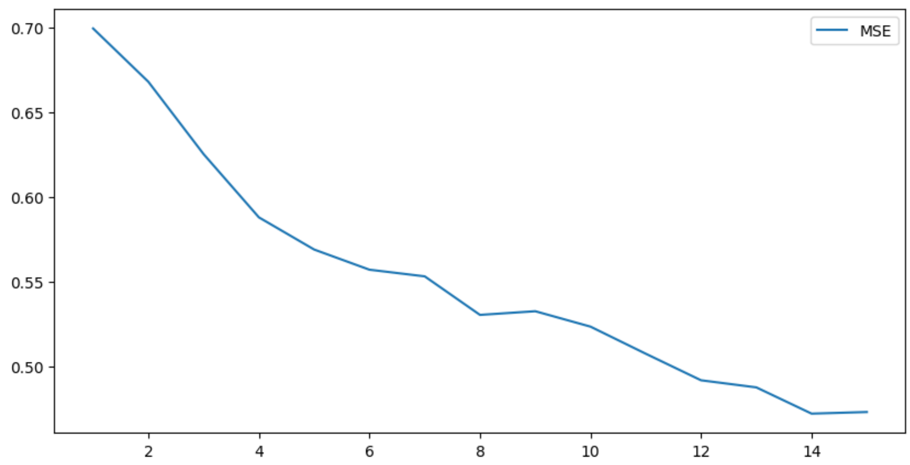

Figure 7: Loss of the training of the LSTM

4.2 Land and ocean tempearture anomaly prediction

Also in this case the original dataset has been splitted with the 80% of the data used for training and the other 20% used for test. The RNN has been trained until convergence of the MSE loss, Figure 7 shows the loss distribution during the training phase.

The test loss achieved by the final model is 0.07, which is so misleading, as it is showed in Figure 8 for the land anomaly tempearture prediction and in Figure 9 for the ocean anomaly temperature prediction.

The LSTM seems to recognize the general trend, but it can not accurately tell the quantity of the anomaly in the temperature.

4.3 Sea level prediction

Also in this case the original dataset has been splitted with the 80% of the data used for training and the other 20% used for test. The RNN has been trained until convergence of the MSE loss, Figure 10 shows the loss distribution during the training phase.

The test loss achieved by the final model is 6.73, and the model predictions can be seen from Figure 11.

Also in this case the LSTM recognize the general pattern but has difficulty on precisely predict the future GMSL value, although in general it approximates decently the true distribution.

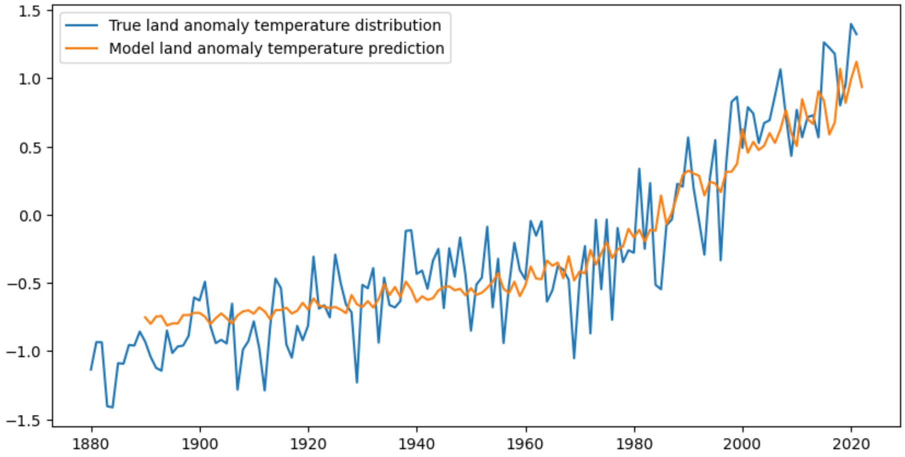

Figure 8: LSTM model predictions on future land anomaly temperature

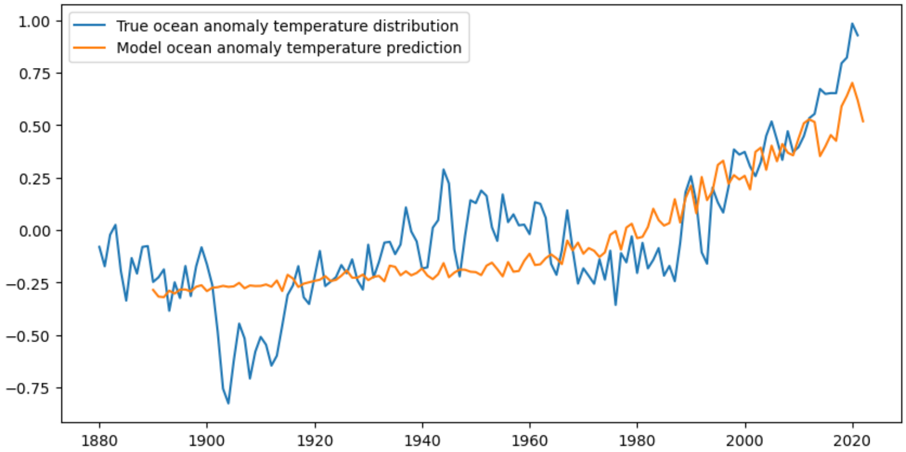

Figure 9: LSTM model predictions on future ocean anomaly temperature

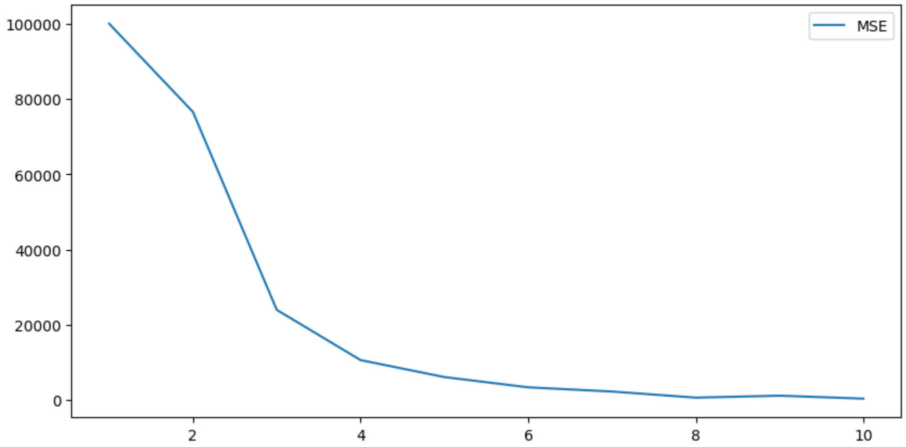

Figure 10: Loss of the training of the LSTM

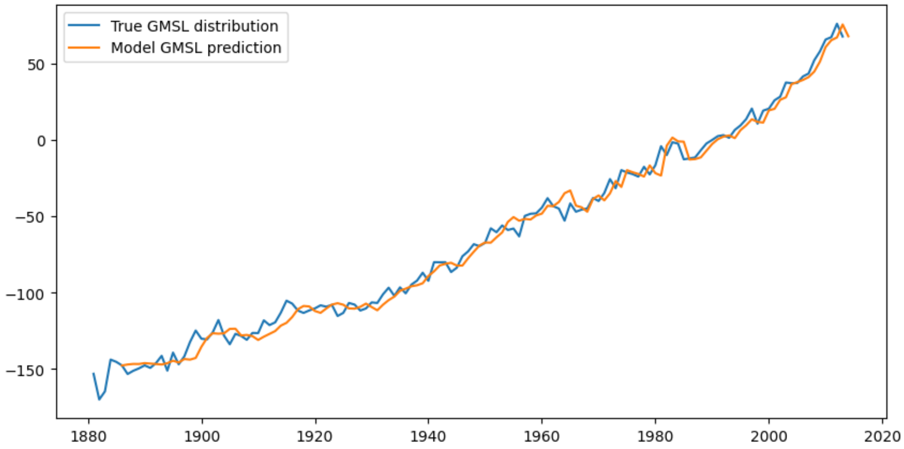

Figure 11: LSTM model predictions on future GMSL values

5 Failure cases

The initial idea of the project was to also predict the *mortality* of the humans in USA and UK (countries for which this type of dataset was easily available) from the anomalies of the temperature. But for this task none of the models where able to learn something, this I think because the cause of mortality are various, for example the distribution could have a peak on the start of the 2020 due to the COVID pandemic. Surely this task can be accomplished with a dataset which specifies the cause of the death, and so being able to remove cases which are not strictly related to the rising of the temperatures and all that goes with it.

Anothe failed task prediction has been the prediction of the number of *wildfiers* in the USA from the anomalies of the temperature. Also in this case the reasons are the same for the *mortality* prediction task, a more specific dataset could be helpful, for example one that specify if the wildfire has been provoked by pyromaniac or not.

6 Final considerations

To put in practice the learnt models some predictions about the future land and ocean temperatures anomalies, sea level and CO_2 emissions are made.

Three scenarios will be modelled for the predictions of the next 20 years:

- **Case 1**: the CO_2 emissions of the next 5 years increase at a rate equal to the recent past years.

- **Case 2**: the CO_2 emissions of the nest 5 years are similar to the recent past years.

- **Case 3**: The CO_2 emissions of the next 5 years decrease at a rate inverse to the recen past years.

The Figures 12, 13, 14 and 15 shows respectively the CO_2 emissions, land and ocean anomaly temperatures and the sea level from now to 20 years for the case 1. For the models, the constant rising of the CO_2 emission will rise the land temperature of 1 celsius degrees, this is a trend that effectively it is expected to be reached in the future years. A similar prediction has been made for the ocean anomaly temperature, while for the sea level the prediction does not seems so reliable, this could be caused by an underfitted LSTM, given that even in the training phase it has difficulty to learn, without reaching a good loss value. This could be cause by the fact that the sea level is conditioned by also other factors, for example by a seasonal galcier melting.

For the case 2 the CO_2 has not be plotted as it has equal values to the recent years and do not change among the future 20 years. The Figure 16 and Figure 17 shows respectively the prediction for the land and ocean anomaly tempeartures, the trend seems to be very similar to the one of case 1, even if the CO_2 emission are still the same. While in Figure 18 the future sea level

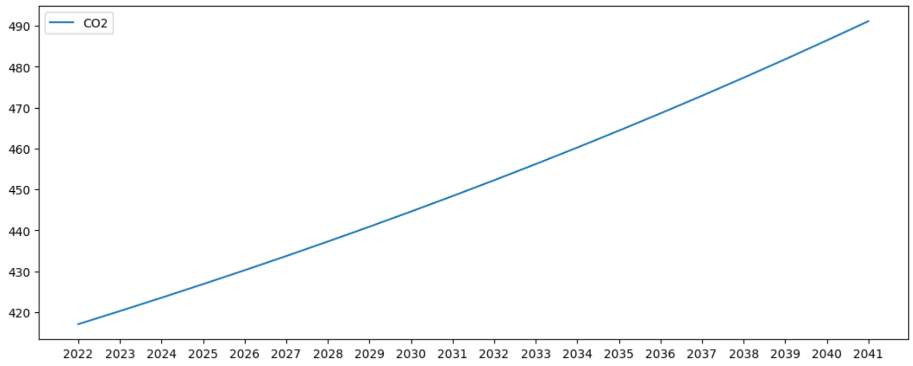

Figure 12: CO_2 future emissions for the case 1

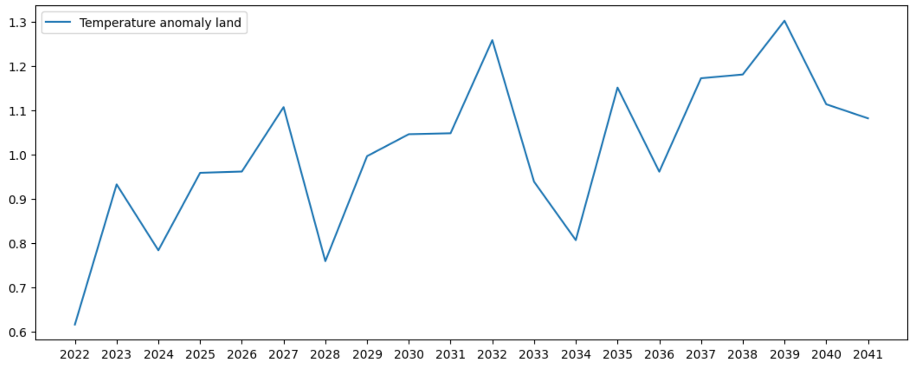

Figure 13: Future land anomaly temperature for the case 1

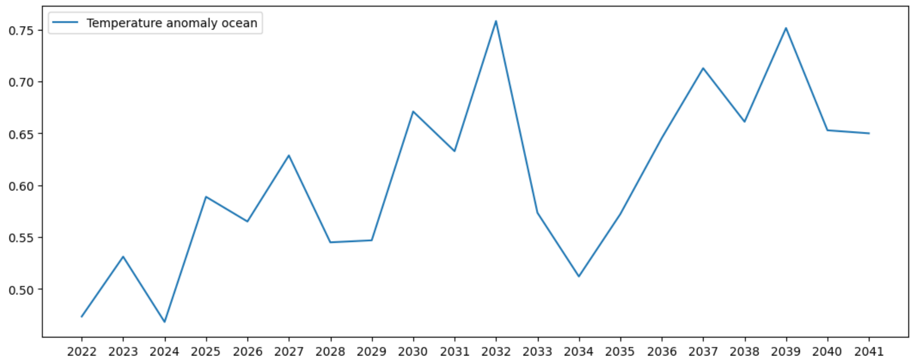

Figure 14: Future ocean anomaly temperature for the case 1

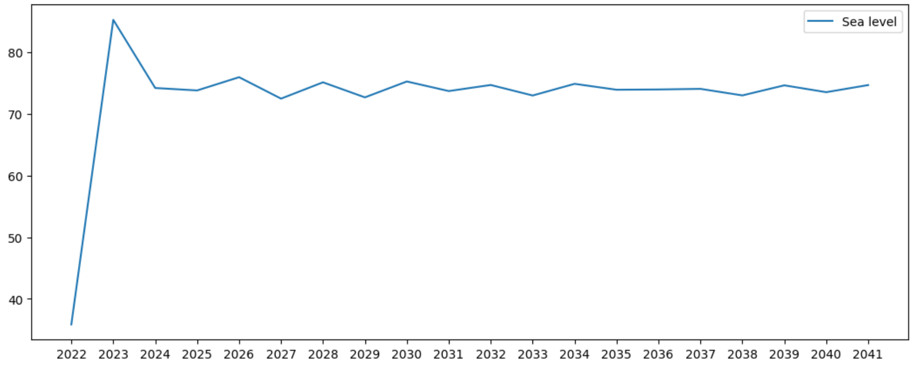

Figure 15: Future sea level for the case 1

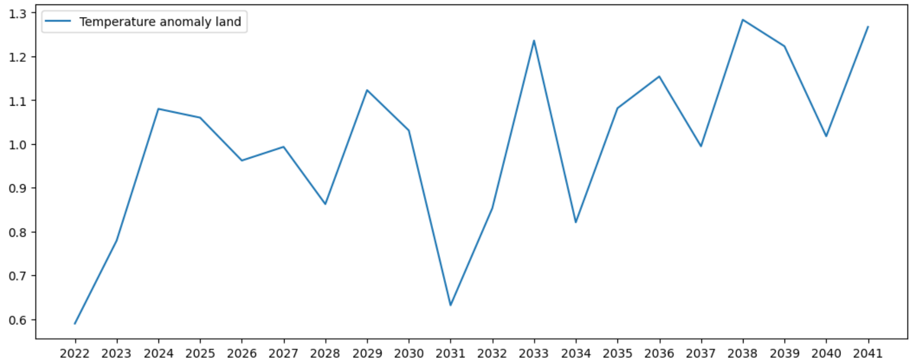

Figure 16: Future land anomaly temperature for the case 2

distribution is showed, also in this case the distribution is similar to the one of case 1, reflecting the poor performances of the trained LSTM model.

The final case is the most promising one, where the CO_2 emissions are low as before the second industrial revolution. Figure 19 and Figure 20 show the distributions of the land and ocean anomaly temperatures. The temperatures are initially low and then they start to increase, this can be a sign of the poor generalization achieved by the LSTM, as this increasing pattern could be classified as bias. In the Figure 21 the sea level prediction model have still difficulty on predictions of unknown instances.

A lot of improvements can be made to reach prediction models that can successfully provide reliable predictions, even if the ones presented in this project report generally provide a good generalization on what is and what will be the scenarios caused by climate change.

Once obtained accurated prediction models, an extention of this project

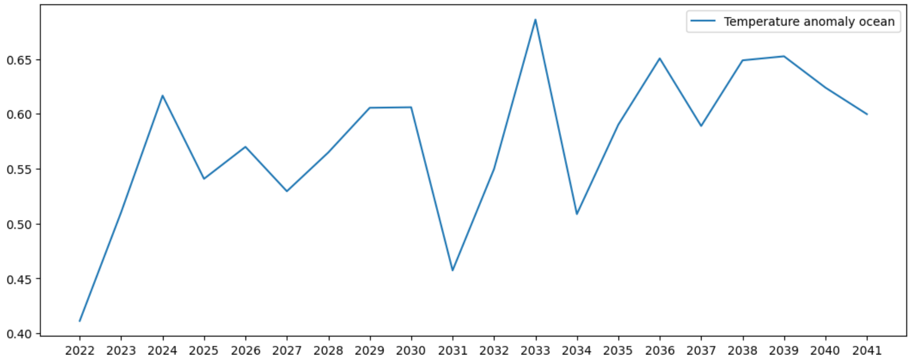

Figure 17: Future ocean anomaly temperature for the case 2

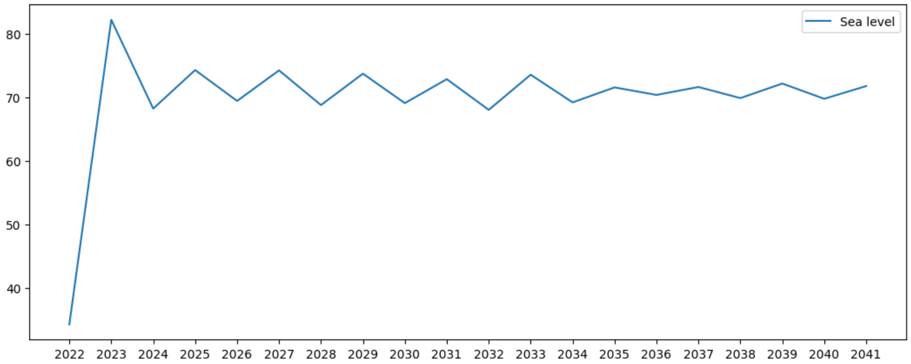

Figure 18: Future sea level for the case 2

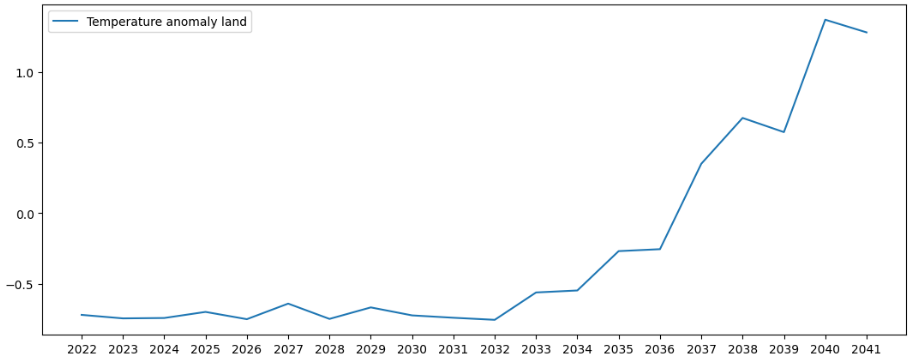

Figure 19: Future land anomaly temperature for the case 3

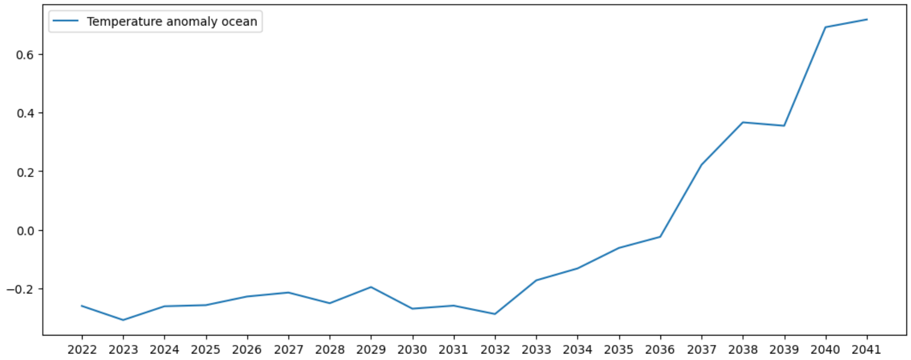

Figure 20: Future ocean anomaly temperature for the case 3

could be to train a *Reinforcement Learning* model to find a policy that will decrease the CO_2 production by taking into consideration the most important causes of the emission of the gas.

References

[1] DM Etheridge, LP Steele, RL Langenfelds, RJ Francey, JM Barnola, and VI Morgan. Historical co2 records from the law dome de08, de08-2, and dss ice cores. *Trends: a compendium of data on global change*, pages 351–364, 1998.

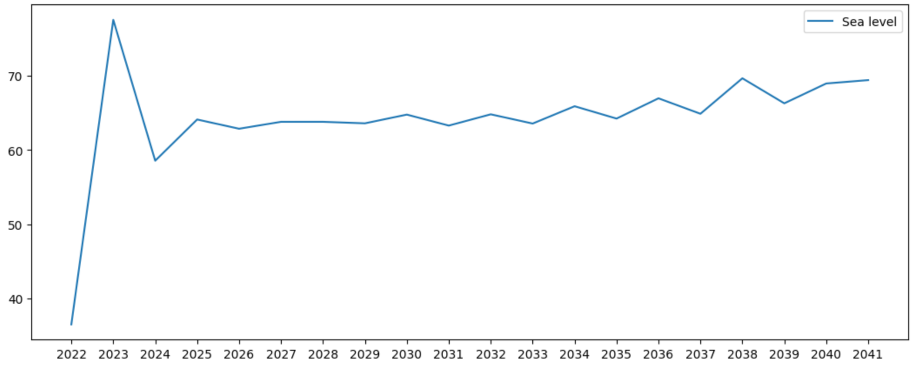

Figure 21: Future sea level for the case 3