Carbon Emission Forecasting
Final Report

Francesco Parisio

February 2, 2022

1 Executive Summary

Background. Precise and accurate forecasts of CO_2 emissions can help policy makers devise strategies and measure their effects in reducing emissions to meet the climate targets. Collected past data can be employed to analyze the problem and data science methods can be used to make forecasts.

Methods. I analyzed existing data from U.S. Energy Information Administration on CO_2 emissions from electricity production per source. After introducing descriptive statistics and a preliminary data analysis, I propose to analyze CO_2 emissions from the largest growing source in electricity production: natural gas. I compare different algorithms (AR, MA, ARMA, ARIMA, SARIMAX) optimized to build a reliable 1 year forecasting model of the non-stationary data. I applied rolling split Monte Carlo cross-validation to evaluate the forecast accuracy.

Results. The optimized SARIMAX model is the best solution that strikes a balance between forecast precision and model complexity. The Symmetric Mean Absolute Percentage Error (SMAPE) on testing data is below 3%. The cross-validation provides a quantification of uncertainty in the forecast, showing an (expected) increasing average error with increasing time horizon of the forecast. The average error remains below 11% and roughly 75% of the forecasts are likely to have a SMAPE< 15%–a proof of the model's reliability and robustness.

Conclusions. The optimized SARIMAX model forecasts CO_2 emissions with a high precision. The forecast shows an increase in CO_2 emissions from natural gas in the coming 12 months that can serve as a basis case to measure policy effects. Future studies should consider multivariate time series analysis that includes relevant indicators of CO_2 emissions (e.g., energy demand, natural gas prices, carbon-bonds prices).

2 Problem and solution summary

2.1 Problem statement

Carbon dioxide (CO_2) is the main contributor to planet Earth's greenhouse effect and its emissions are one of the key indicators of the anthropogenic impact on climate change [6]. National climate policies and international agreements are needed to reach the climate targets of net zero emissions and climate neutrality by mid-century and contain the global temperature rise within 1.5-2 °C [12].

To devise proper policies and actions, we need a basic understanding of the main contributors to CO_2 emissions and how emissions evolve with time. The capability to forecast emissions of CO_2 is relevant in order to build future climate scenarios; containing the global temperature rise within 2°C means that each individual country will have to adjust its climate policy and measure the success of the action based on future trend comparison with the available forecasts. Better forecasts mean a higher efficacy of climate policies, better targeted policies and more effective public interventions.

The production of electricity plays a crucial role in the U.S. as a major source of CO_2 emissions [11]. Reducing CO_2 emissions without an increase in the cost of electricity is possible [9]. The introduction of a carbon-tax that affects particularly polluting sectors [13], the shift toward carbon-free renewable energy sources [3] and the development of carbon capture and storage technologies [4] are all potential policies that can contribute to CO_2 emissions reduction of the electric energy sector without excessively affecting its costs.

Accurate forecasts are used to make policy decisions and to measure their efficacy by comparing forecasts with future records. In this sense, forecasts accuracy is fundamental for climate policy: the goal

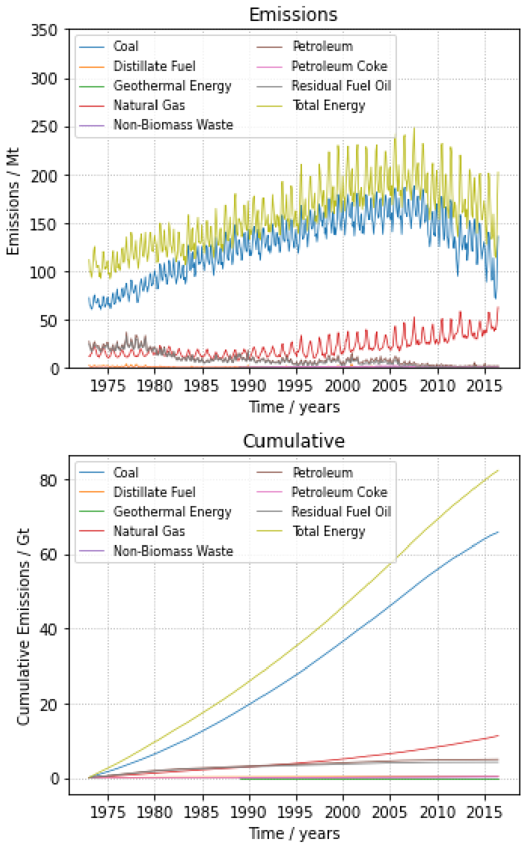

Figure 1: Evolution in time and cumulative values of CO_2 emissions from all energy sources.

of this project is to build a forecasting methodology based on available data. I will analyze data about CO_2 emissions for electricity production in the U.S. from 1973 to 2016. The objectives are further split into sub-categories. The first part presents preliminary descriptive statistics and visual analysis to get a deeper understanding of the problem by extracting useful information from historical records. The second part explores different methodologies to forecast the CO_2 emissions from natural gas for the next 12 months period. Finally, the outcome of the forecast is presented along with implementation strategy and recommendations for future actions.

2.2 Solution design

2.2.1 Preliminary data analysis and visualization

The initial data contain 384 null values in time and, after cleanup, totals 4323 entries. By grouping the emissions from each different energy source, it is possible to plot the individual contribution of type of source of electricity production to the total emissions (Figure 1). Electricity generated from coal has always been the largest CO_2 emitter and the amount is almost equivalent to the total. Natural gas is the second emitter in recent years, a rank occupied by residual oil and petroleum until roughly the early 1980s. From approximately 2005, petroleum and residual oil emit negligible amounts, as do all emissions other than natural gas and coal. While a continuous increase has been observed in total emissions from the 1970s to 2005, emissions started to decrease in recent times. Some of the energy types are not visible because emissions are too small compared to the total. The cumulative emissions over time shows that electric energy production from coal has been the biggest source of CO_2 emissions in the past. Natural gas, petroleum and residual fuel oil are only a fraction of it and the other types are practically negligible. While coal is the largest emitter, its decline brings hope that it will slowly be phased out as an electricity generation source, even in the light of recent policies aimed at revitalizing the industry [10]. On the contrary, the steady increase in emissions of the last three decades from electricity produced from natural gas, the second emitter, should be a concern for policy makers that aim at reducing emissions. Therefore, we will focus the analysis on emissions from natural gas.

2.2.2 Stationarity tests and decomposition

The first step in the analysis consists in splitting the natural gas emissions in training and validation: the split preserves the time order and the training is from beginning until 2015-08-31, while the testing is from 2015-08-31 to 2016-07-31. After splitting the data, the next step is to analyze whether or not it is stationary. This can be analyzed graphically, by plotting the rolling mean on a 12 windows observation (MVA12) and by taking a log-shift transformation $\bar{x}_t = \log x_t - \log x_{t-1}$. Figure 2 shows that the original data is non stationary because the MVA12 gradually increases with time, showing a clear trend in the data. On the contrary, the MVA12 applied to the log-shift transformation oscillates around the null value, which implies that the log-shift transformed series is stationary. This is further confirmed by the augmented Dickey–Fuller test (ADF) assuming a critical p-value of $p_{\text{crit}} = 0.05$. The test yields a p-value of $0.997 > p_{\text{crit}}$ for the original series and a p-value of $3.84 \times 10^{-5} < p_{\text{crit}}$ for the log-shifted series. The ADF test confirms that the original series is non-stationarity, while the log-shifted series is stationarity.

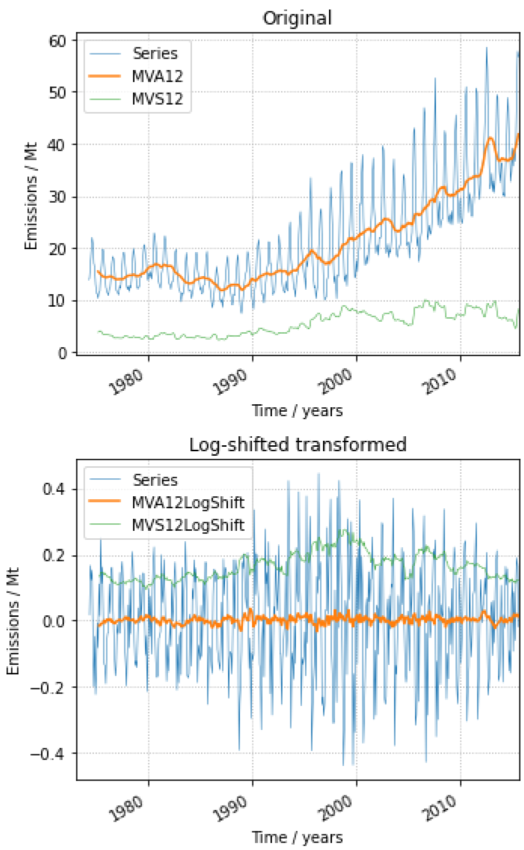

Figure 2: Visual test of stationarity of the original and log-shifted data series.

Decomposing the time series is useful to study the evolution of trend and seasonality. Figure 3 shows an increasing trend since the year ≈ 1990. The seasonality zoom shows an increment in emissions in the summer months that is likely related to the seasonality of electricity demand in the U.S. as a consequence of the air conditioning needs. It has been established that climate change could increase demand for residential air conditioning [7]: this fact could be particularly relevant as it might trigger a feedback loop where emissions increase the temperature, which increases cooling demands, which increases emissions back. The MVA12 of the residuals is always ≈ 0, which means that removing the trend and seasonality to the time series makes it stationary.

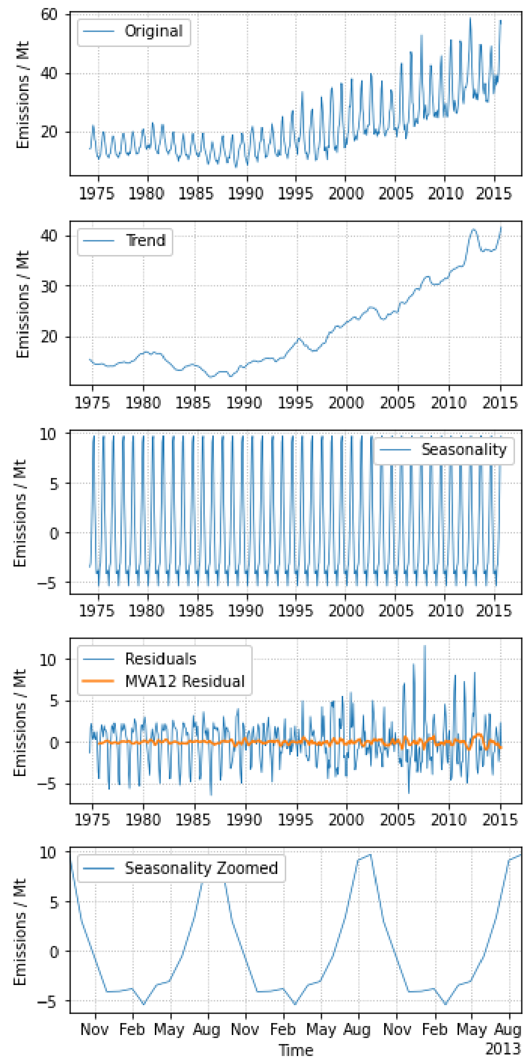

Figure 3: Decomposition of the time series in trend, seasonality, residual and a further zoom into the seasonal component.

2.2.3 Model search

In order to forecast the CO_2 emissions, a proper model must be selected. The candidates for the model selection are Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX). The parameters of the models are listed in Table 1.

The plots of the auto-correlation function (ACF) and partial auto-correlation function (PACF) can be of guidance to estimate the optimal parameters (Figure 4). PCF shows that after a lag $p = 1$, it becomes

Table 1: Forecasting models compared with parameters values, AIC score and SMAPE on validation data.

Model	Parameters	AIC	SMAPE
Non-optimized models (lag from PACF)			
AR	p=1	-480	24.20%
MA	q=1	-470	24.46%
ARMA	p=1, q=1	-480	24.10%
ARIMA	p=1, d=1, q=1	-475	24.13%
Models optimized through grid-search			
Optimized ARIMA	p=12, d=2, q=4	2190	2.64%
Optimized SARIMAX	p=2, d=1, q=0, P=0, D=1, Q=1, s=12	1999	2.84%

statistically insignificant, but then increases again. Consequently, for the AR component, we will assume $p = 1$. The spike at $p = 12$ is likely indicating the seasonality component of the emissions which occurs at a frequency of 12 months (yearly). The remaining parameters for ARIMA are also set to unity, i.e., $q = 1$ and $d = 1$. The models are fitted on the training data and the Symmetric Mean Absolute Percentage Error (SMAPE) is employed to estimate the goodness of the fit on the testing data. Furthermore, the Akaike Information Criterion (AIC) is employed to contribute in the evaluation of the best model by providing a balanced estimate between under and over-fitting. The results are summarized in Table 1. AR, MA, ARMA and ARIMA models perform very poorly on the validation data, with SMAPE$>$ 24%, most likely because seasonality of the time series is not correctly captured. The AIC for the non-optimized models are very similar, with the best value for the ARMA and AR models. The non-optimized model are trained on the log-shifted data (stationary), but do not perform well enough on the testing data.

To overcome the problem under-fitting, two models (ARIMA and SARIMAX) have been optimized by performing a grid search on the input parameters and by measuring RMSE on the validation data. The optimized ARIMA model shows an drastic improve in performance, with SMAPE$=$ 2.64%. The AIC has drastically increased to AIC$=$ 2190, since the model is much more complex (more coefficients) than the un-optimized one. The optimized SARIMAX model also shows very good performance in the testing data with SMAPE$=$ 2.84% that is very close to the one of optimized ARIMA.

The comparison between the forecast from the optimized models against the validation data is shown

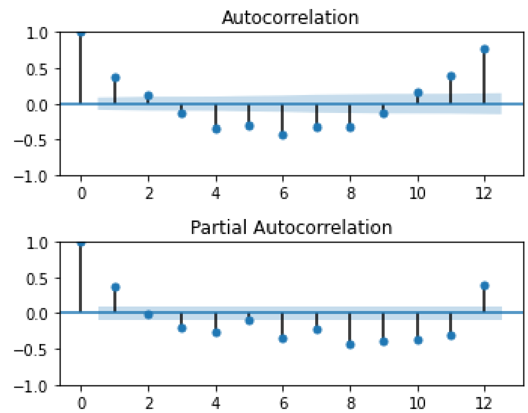

Figure 4: ACF and PACF plots can provide estimates of the model parameters.

in Figure 5. Both models show a very good fit against the validation data and have confidence intervals (light shaded area) that contain the validation data. However, since the AIC score is lower for the optimized SARIMAX model than for the optimized ARIMA, the optimized SARIMAX model is the final choice for the forecast.

2.3 The SARIMAX model

2.3.1 Monte Carlo rolling split cross-validation

In this section, I present a Monte Carlo strategy for a rolling split cross-validation of the optimized SARIMAX model. The cross-validation of time series implies that the series has to be split by preserving the time order. The algorithm splits the series at a random time in the last 30 years ($t_0 = 1986 - 2016$) of the data series. At the split time, the past is considered as the training data and the future (12 months ahead) as the testing data. At each run, a

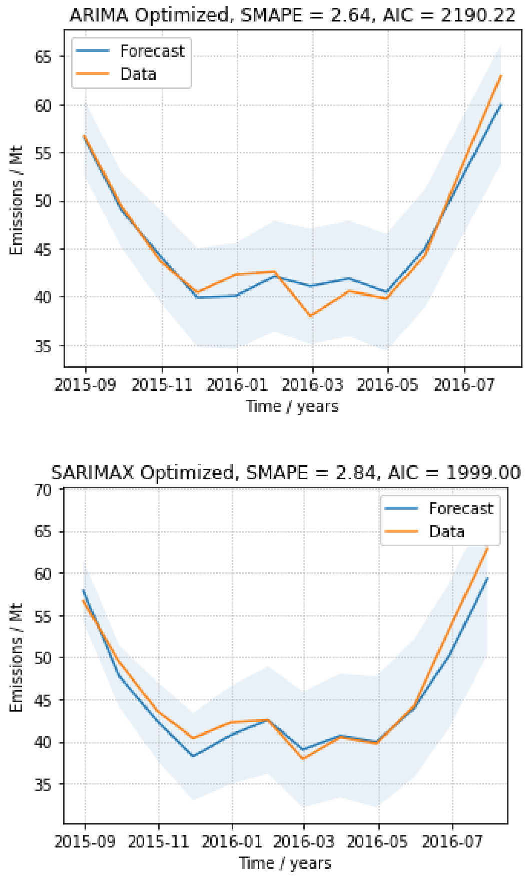

Figure 5: Comparison with validation data of the optimized ARIMA and SARIMAX forecast, with confidence intervals in light shaded color.

random split is assigned in terms of the splitting time t_0, the model is trained and the forecast performance is compared with observations on the testing data. The performance is evaluated in terms of SMAPE between forecasts. All data is aggregated and for each month ahead of the forecast, the average SMAPE of all simulations is computed. A total of 10000 simulations is carried out and the confidence interval of the SMAPE is computed assuming a p-value of 0.95. Finally, for each simulation, overall 1-year ahead SMAPE is computed and the results are aggregated in the final statistics.

Figure 6 shows the time evolution ahead of the forecast time t_0 of the average SMAPE computed from all simulations. The forecast quality is worsening in the first 8 months after time t_0 (increasing SMAPE) and becomes approximately constant around $\approx 11\%$ in the range 8 to 12 months of forecast. This result indicates that prediction quality decreases with increasing time ahead of the forecast (model drift). On average, SMAPE falls between 7% to 11% roughly with a slightly increasing confidence interval with time. The latter implies an increase in uncertainty in the forecasting power with the time ahead of t_0.

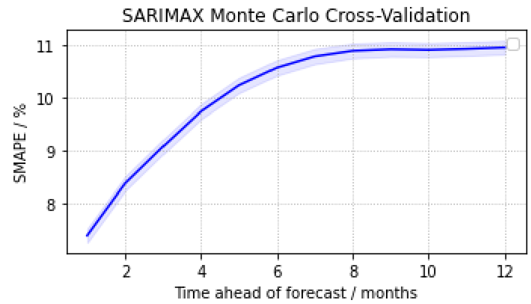

Figure 6: Time evolution of the average from all 10000 simulations of the normalized absolute residuals between forecast and data.

Figure 7 shows the summarizing statistics of the SMAPE distribution in terms of boxplot. The median SMAPE is around 10% and most simulations fall within 15% of SMAPE. Outliers of very high SMAPE ($> 25\,\%$) are present but are quite unlikely ($2.46\,\%$). The Monte Carlo approach to rolling split cross-validation shows that if the optimized SARIMAX model would have been applied to perform 1 year forecasts in the last 30 years, it would have performed 50% of the times with a SMAPE lower or equal to 10%. Additionally, on average the forecast would have shown a decrease reliability with time ahead of the forecast, but with an average SMAPE on the prediction that remains below 11%. We can conclude that the model can provide highly reliable and robust forecasts.

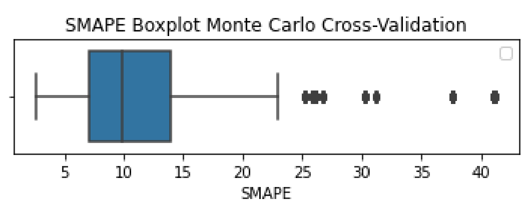

Figure 7: Descriptive statistics of the SMAPE distribution summarized in a boxplot.

2.3.2 Forecast of 1 year ahead

The optimized SARIMAX model presented in the previous section is employed to generate forecasts

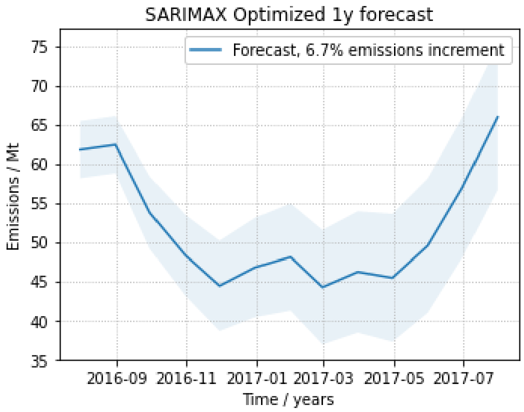

Figure 8: 1 year forecast of CO_2 emissions from electricity production with natural gas. Light colored shade indicate the confidence interval of the forecast (increasing with time).

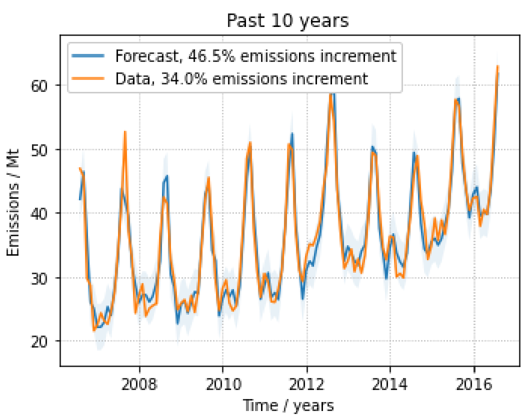

Figure 9: Original data and model prediction in the past 10 years used to compare average yearly increment.

for 1 year ahead of the maximum time of the data. Figure 8 shows the 1 year forecast along with the confidence interval in light-shaded color. The forecast shows a decrease in emission in the first semester, followed by an increase in the second semester. The oscillation follows the seasonality of the past observations, where emissions are higher in the summer months. The forecast value shows an increment in CO_2 emissions of 6.7% for 2017.

To better understand how this value of increment in the forecast emissions compares with historical data, Figure 9 shows the emission and the model fit on the 10 years prior to t_0 time of the forecast. Data are re-sampled each year on July 31st for consistency. In the prior 10 years, an average of 4.6% increment per-year has been observed in the model, compared to an average 3.4% increment per-year from the recorded data. This implies that the model forecasts for 2017 an increase in CO_2 annual emissions that is greater than the average annual increment observed (or modeled) in the prior decade. In other words, the situation is not only worsening (more CO_2 emissions from natural gas for electricity production), but the model indicates for 2017 a situation worsening at an increasing pace.

3 Recommendations for implementation

3.1 Recommendations for implementation

The optimized SARIMAX model has shown great capabilities to forecast CO_2 emissions for electricity production from natural gas. The model can be easily and quickly deployed as a forecasting tool and can be launched as a continuously integrated background forecasting model. In the deployment, the training and validation is very rapid, and so is the grid search optimization implemented. Since new data will become available each month, further re-optimizing and re-training should be carried out. The optimization of the SARIMAX model can be improved in terms of time performance. To speed-up the problem, the loops could be further optimized through parallelization. The bottleneck in terms of computing performance is the Monte Carlo cross-validation, which takes approximately 50 minutes to run on a single GPU. However, if data become available monthly, performing a re-optimization, re-training and re-run of the Monte Carlo algorithm would be quite simple to implement and fit in reasonable time-frame.

3.2 Actions for stakeholders

The optimized SARIMAX model that I propose is a valuable tool for policy makers. It can be used to forecast CO_2 emissions that form a baseline value. The baseline can be employed to measure effectiveness of given policy choices. Through the Monte Carlo approach, it was shown how the model performance is very high and that the average error on the 1 year forecast is approximately 11% and decreases to 7% for forecasts of 1 month horizon.

Continuous integration and model maintenance ensures that forecast quality will not degrade in time and will remain at acceptable level. With a simple procedure, it will be possible to derive a baseline for more precise and targeted actions and policies. For example, if the models forecasts an increase of 7%, a 20% reduction of the emissions from the forecast could become the target, limiting emissions increase to 5.4%. However, as new data becomes available monthly, the forecast will be updated and can provide a rationale to decide whether or not the current actions are driving emissions towards the annual target. In this way, the optimized SARIMAX can strongly contribute to reach the climate targets.

3.3 Benefits vs costs

There are costs that concern model implementation itself, and other are connected to intrinsic cost of the specific policies adopted. The benefits are quite clear: improved climate, air quality and a step closer to meet the climate targets.

The optimized SARIMAX method is a tool to be evaluate policy efficacy. It can be of great benefit in deciding policy actions and in measuring their efficacy. Some maintenance of the forecasting model is to be expected, but it will be minimal given the high efficiency of training the model and performing the forecast. If new data will become available on a monthly basis, I expect a maintenance cost that will not exceed 8 person-hour per month.

The additional costs that the policy makers have to face are a consequence of promoting a green energy alternatives. While specific road-maps have been proposed for the energy transition towards a mainly hydro, wind and solar based production [8], the action can be accelerated by setting yearly targets based on accurate forecasts for each sector. Thanks to a reliable forecasting model, the costs can be minimized by targeting the action policies (carbon tax, private incentives, research spending) that provide the highest benefits. Reduced CO_2 emissions improve the health of citizens, reducing the burden on the healthcare system, improve climate, reduce planetary greenhouse effect and promote the growth of green energy sector that is necessary to meet climate targets. Globally, the ratio costs vs benefits is very low, because the great advantages that policy makers can draw from reliable forecasts come at the very low cost of operating and maintaining the forecasting platform through time-based updates.

3.4 Risks, challenges and mitigation strategies

One of the major risk of the proposed solution is characteristic of any time series forecast: the inability to forecast black swan events. Although the historical time series of CO_2 emissions for electricity production from natural gas sources did not show black swan events in the past, those cannot be excluded in the future. This has to be taken into proper account when deploying the model to make forecasts and can be mitigate by incorporating expert and domain knowledge into the forecast.

Another challenge is linked to the degrading forecast quality with increasing time horizon (model drift). This is an intrinsic characteristic of time series forecast, since those are extrapolation problems in contrast with, e.g., regression problems that are usually applied to interpolate. The further the extrapolation from the training domain, the higher is the error. The assumption is demonstrated and quantified in this report by the Monte Carlo analysis. Nonetheless, if forecasts remain within 1 year ahead, the chances of errors greater than 25% is lower than 3%.

3.5 Outlook and perspectives

In this project, univariate time series analysis was applied. From the analyses carried out, it was shown that an optimized SARIMAX model can provide highly reliable forecast over 1 year ahead. The forecast has shown not only an increase in the CO_2 emissions in the coming year, but also an increase which occurs at a higher rate than the average increment in the previous 10 years. This finding is a clear indication for policy makers: strong actions should be deployed as quickly as possible to reduce CO_2 emissions. Furthermore, the emissions peak in the summer months, most likely linked with the cooling needs. With increasing global temperatures and the increasing likelihood of extreme events such as prolonged and intense heat waves, reducing CO_2 emissions might become increasingly difficult in the years to come. This brings an additional urgency upon policy makers to rapidly devise and implement appropriate actions to reduce CO_2 emissions.

In order to get a deeper and broader understanding of the problem, other indicators should be employed in the data. Adding new indicators would make it possible to carry out time series forecasts from multivariate analyses. More complex methods can be used in multivariate time series analyses, such as the Long-Short Term Memory [5] and the multivariate Singular Spectrum Analysis [2]. Additionally, by analyzing emission data from other countries, it would be possible to employ the Synthetic Control Method [1] to evaluate most effective past policies in reducing CO_2 emissions.

References

[1] A. Abadie, A. Diamond, and J. Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

[2] A. Agarwal, A. Alomar, and D. Shah. On multivariate singular spectrum analysis and its variants. *arXiv preprint arXiv:2006.13448*, 2020.

[3] E. Dogan and I. Ozturk. The influence of renewable and non-renewable energy consumption and real income on co2 emissions in the usa: evidence from structural break tests. *Environmental Science and Pollution Research*, 24:10846–10854, 2017.

[4] J. D. Figueroa, T. E. Fout, S. I. Plasynski, H. G. Mcilvried, and R. D. Srivastava. Advances in co2 capture technology—the u.s. department of energy's carbon sequestration program. *International Journal of Greenhouse Gas Control*, 2:9–20, 2008.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] D. Holtz-Eakin and T. M. Selden. Stoking the fires? co2 emissions and economic growth. *Journal of public economics*, 57(1):85–101, 1995.

[7] M. Isaac and D. P. van Vuuren. Modeling global residential sector energy demand for heating and air conditioning in the context of climate change. *Energy Policy*, 37:507–521, 2009.

[8] M. Z. Jacobson, M. A. Delucchi, G. Bazouin, Z. A. Bauer, C. C. Heavey, E. Fisher, S. B. Morris, D. J. Piekutowski, T. A. Vencill, and T. W. Yeskoo. 100% clean and renewable wind, water, and sunlight (wws) all-sector energy roadmaps for the 50 united states. *Energy & Environmental Science*, 8(7):2093–2117, 2015.

[9] A. E. MacDonald, C. T. M. Clack, A. Alexander, A. D. Dunbar, J. M. Wilczak, and Y. Xie. Future cost-competitive electricity systems and their impact on us co2 emissions. *Nature Climate Change*, 6:526–531, 2016.

[10] R. Mendelevitch, C. Hauenstein, and F. Holz. The death spiral of coal in the us: will changes in us policy turn the tide? *Climate Policy*, 19(10):1310–1324, 2019.

[11] X. ting Jiang and R. Li. Decoupling and decomposition analysis of carbon emissions from electric output in the united states. *Sustainability*, 9:886, 2017.

[12] J. Tollefson et al. Cop26 climate summit: A scientists' guide to a momentous meeting. *Nature*, 599(7883):15–17, 2021.

[13] I. S. Wing. The synthesis of bottom-up and top-down approaches to climate policy modeling : Electric power technologies and the cost of limiting us co2 emissions. *Energy Policy*, 34:3847–3869, 2005.