Predicting Temperature with Monthly Global Average Methane Concentrations

Jingze Dai, Rachael Stephan, & Zhuocheng Zhang

April 25, 2025

Introduction

Greenhouse gases (GHG) are natural, atmospheric gases that can absorb infrared radiation[1]. GHGs will reabsorb heat energy emitted by Earth and retain it within the atmosphere, re-radiating heat back towards the surface[1]. This phenomena is known as the greenhouse effect. Therefore, GHGs are important determinants of Earth's energy budget and resulting climate[1]. Concentrations of greenhouse gases will naturally oscillate and there are other external climate drivers as well (e.g., Milankovich cycles)[1]. However, the anthropogenic consumption of fossil fuels since the Industrial Revolution have resulted in elevated concentrations of GHGs[1]. The most influential greenhouse gases with regards to contribution to the greenhouse effect are carbon dioxide, methane, and water[1].

Methane is commonly regarded as the second most important GHG compared to carbon dioxide - especially considering methane has lower atmospheric concentrations, but carbon dioxide and methane are difficult to compare due to their residence time and radiative forcing capabilities[2]. Methane has a more potent radiative forcing (i.e., higher heat retention) but a residence time on the order of decades compared to carbon dioxide's centuries[1]. Assuming an equal amount of each gas is released, this will result in methane having more power over climate warming than carbon dioxide at small time scales[2]. Over 20 years in this scenario, methane will be 80 times as powerful as carbon dioxide[2]. Over the next 100 years in this scenario, methane will be 28 times as powerful as carbon dioxide[2]. Therefore, methane is an important consideration for near-future climate considerations, such as within our lifetime.

Two of the major contributors to methane emissions are agricultural practices and fossil fuels[3]. The fermentation during the digestion process in cattle is the primary source of agricultural methane. It is estimated that up to 264 pounds of methane worldwide are emitted from cattle annually[4]. Methane emitted through fossil fuels are caused by extraction leaks, crude oil processing, and transportation of the fossil fuels. This results in an estimated global annual emission of 120 million metric tons[5].

Monitoring atmospheric GHG concentrations is important for climate monitoring and tracking the effectiveness of climate change mitigation – including in the green energy transition. This report will investigate the temporal trends in methane concentrations considering the exogenous factors of beef production and socioeconomic factors (as a proxy for fossil fuel consumption). We will use our methane projections to extrapolate trends in climate, represented by temperature anomalies based on an averaged climate from 1901 to 2000.

[1] Mann, M.E. (2025, April 11). greenhouse gas. Encyclopedia Britannica. https://www.britannica.com/science/greenhouse-gas

[2] MIT Climate (2024, January 4). Why do we compare methane to carbon dioxide over a 100-year timeframe? Are we underrating the importance of methane emissions? MIT Climate Portal. tps://climate.mit.edu/ask-mit/why-do-we-compare-methane-carbon-dioxide-over-100-year-timeframe-are-we-underrating

[3] Raymond, P., & Hamburg, S. (2024, November 18). Yale experts explain methane emissions. Yale Sustainability. https://sustainability.yale.edu/explainers/yale-experts-explain-methane-emissions

[4] U.S. Environmental Protection Agency. (2020, October). Agriculture and aquaculture: Food for thought. https://www.epa.gov/snep/agriculture-and-aquaculture-food-thought

[5] International Energy Agency. (2024). Global Methane Tracker 2024: Key findings. IEA. https://www.iea.org/reports/global-methane-tracker-2024/key-findings

Data Sources

The following data sources were used in this analysis and their respective wrangling is described below.

1. *Globally averaged methane concentration data*

- Frequency: monthly
- Units: ppb
- Format: csv file
- Source: NOAA Global Monitoring Laboratory

The initial methane data set was an excel `.csv` file with columns of the year, month, decimal year, average methane concentration, average uncertainty, trend value, and trend uncertainty on a single sheet. It was processed as follows:

- The header and data were read in as separate data frames because of the excel file formatting.
- The header information was set as the column names for the data.
- A date was created using the month and year columns using lubridate's `make_date` function. The day was assumed to be the first of the month.
- Data set was verified for missing data - none were present.
- Data frames were saved as the full data, training period data, and test period data. The training and testing periods are expanded upon in the methods section.

2. *Globally averaged temperature anomaly data (based on a 1901-2000 average)*

- Frequency: monthly
- Units: degrees Celsius
- Format: csv file
- Source: NOAA National Center for Environmental Information

The initial temperature anomaly data set was an excel `.csv` file with columns of the date and temperature anomaly on a single sheet. It was processed as follows:

- Only the data was read into a data frame from the excel file.
- The column names for the data frame were manually set.
- The date column was converting into date format using lubridate's `ym` function.
- Data set was verified for missing data - none were present.
- Data frames were saved as the full data, training period data, and test period data. The training and testing periods are expanded upon in the methods section.

3. *Aggregated global beef production*

- Frequency: yearly
- Units: metric tons
- Format: csv file
- Source: U.S. Department of Agriculture

4. *Socioeconomic factors data set including consumer price index (CPI), industrial production, merchandise exports, and merchandise imports.*

- Frequency: monthly

- Units: index (2005=100)
- Format: csv file
- Source Federal Reserve Bank of Dallas

It was processed as follows: - The date column was converting into date format using lubridate's `ymd` function. - Use the `filter` function to select the dates that match the methane data, and then save to the processed `.csv` file. - There is no missing data, and there is also no outliers.

Data Summary

The data frames were not combined for analysis, but the relevant factors have been merged into a data frame for summary purposes below (Table 1). All of the monthly data is shown in a single data frame (Table 2) and yearly data (i.e., beef production) is shown in another data frame (Table 3).

Table 1: Summary statistics for all of the variables used in the analysis from July 1, 1983 to November 1, 2024, Beef Production is provided until 2023.

Variable	Units	Count	Mean	SD	Min	Max	Pct25	Pct75
Methane	ppb	497	1781.27	70.59	1625.96	1941.81	1739.23	1820.87
Temperature.Anomaly	degrees Celsius	497	0.61	0.29	0.02	1.44	0.39	0.8
CPI	index (2005=100)	497	93.25	57.34	4.35	208.43	45.95	139.02
Exports	index (2005=100)	497	105.18	71.27	14.42	259.8	38.7	173.05
Imports	index (2005=100)	497	82.87	31.05	30.64	140.91	47.9	112.91
Industrial.Production	index (2005=100)	497	99.84	38.8	43.65	177.6	62.6	132.51
Beef Production	metric tons (million)	41	619.72	81.01	490.34412	765.6081	545.37624	692.5579

Table 2: First 10 cases of global methane (ppb), temperature anomaly (degrees celsius), and socioeconomic indicators

Date	Methane	Temperature.Anomaly	CPI	Exports	Imports	Industrial.Production
1983-07-01	1625.96	0.23	4.35	14.42	30.64	43.65
1983-08-01	1628.05	0.36	4.41	14.72	31.78	44.29
1983-09-01	1638.43	0.42	4.47	15.05	31.32	44.77
1983-10-01	1644.80	0.22	4.53	15.00	33.81	44.78
1983-11-01	1642.61	0.35	4.59	15.23	32.02	45.15
1983-12-01	1639.54	0.24	4.65	15.28	31.74	45.61
1984-01-01	1638.73	0.31	4.71	15.80	35.95	45.97
1984-02-01	1638.82	0.20	4.77	16.00	36.07	46.24
1984-03-01	1640.87	0.31	4.84	16.03	36.68	46.20
1984-04-01	1643.96	0.12	4.90	15.66	38.07	46.17

Table 3: First 10 cases of beef production (metric tons)

Year	Beef.Production
1983	490.3441
1984	504.3934
1985	513.5100

Year	Beef.Production
1986	531.0788
1987	530.7524
1988	535.6858
1989	537.9553
1990	553.5117
1991	553.5116
1992	545.3762

All of the identified variables have a relatively strong correlation with each other (Fig. 1). The highest correlation between exogenous variables and methane is CPI with 0.97. The correlation between temperature and methane is 0.87.

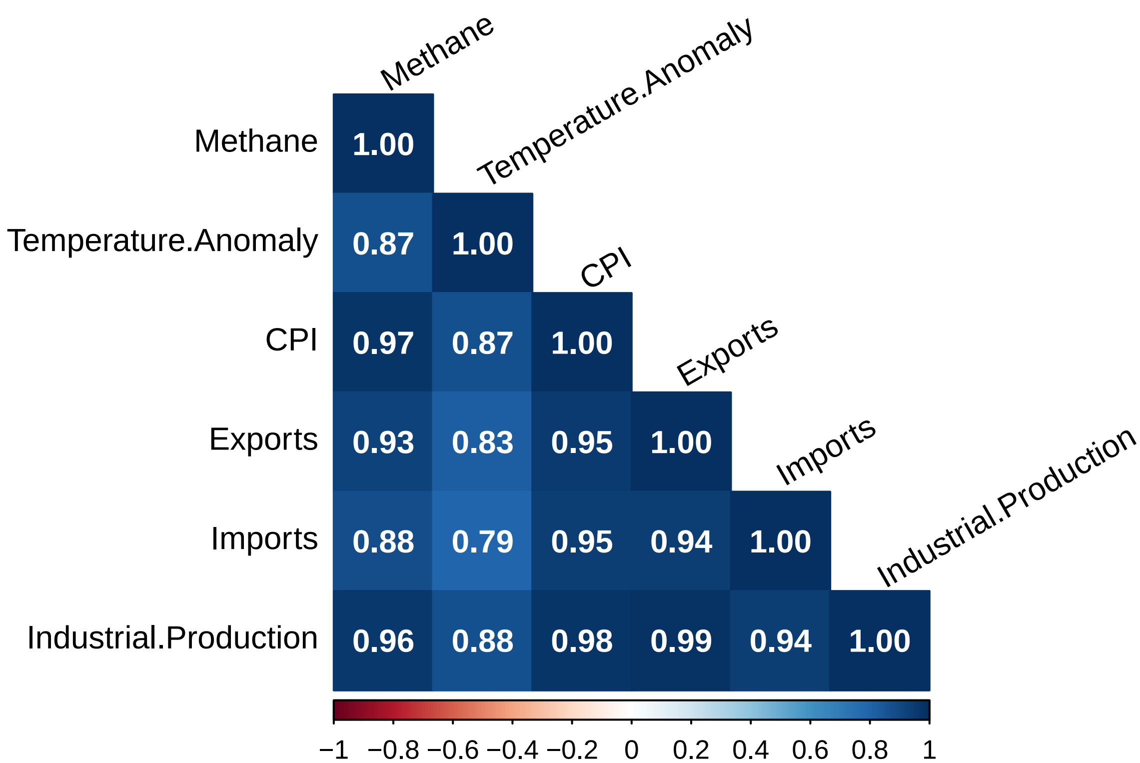


Figure 1: Correlation Between All Monthly Variables

Analysis

Libraries

The following libraries were used in the analysis:

- `tidyverse`: a collection of R packages with similar grammar data structures used to tidy and process data, including `ggplot2` and `lubridate` which are for graphing and handling dates respectively.

- `forecast`: methods and tools used to produce and analyze univariate time series forecasts
- `cowplot`: an addition to `ggplot2` used to align figures
- `Kendall`: contains the Kendall and Mann-Kendall statistical methods
- `tseries`: tools for time series analysis and computational finance
- `corrplot`: a visual exploratory tool for plotting correlation matrices
- `smooth`: tools for state space modelling

Training and Test Periods

The training period was designated as July 1, 1983 to December 1, 2021. The test period was designated as January 1, 2022 to November 1, 2024. These periods were used to filter the data to create test and training data subsets.

Methane Concentration Predictions

The methane data has an increasing trend with a seasonal component (Fig. 2).

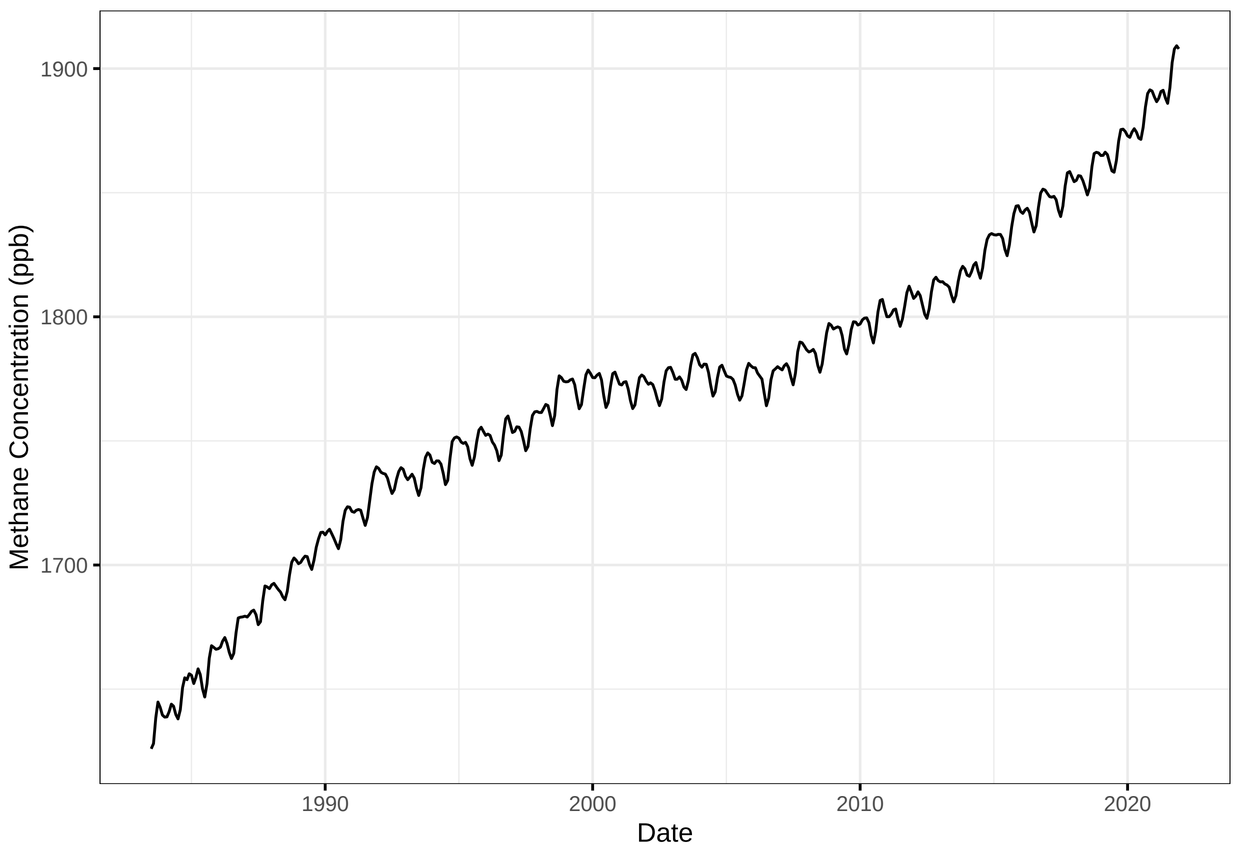


Figure 2: Globally Averaged Methane Concentrations From July 1, 1983 - November 1, 2024

Decomposing the methane time series confirms a positive trend with a seasonal component (Fig. 3). The random component is relatively evenly distributed, but there may be some seasonality or other periodic component remaining because the peaks and troughs appear to be fairly regular.

The ACF plot shows a gradual decay and the PACF plot shows a sharp drop off after lag 1 (Fig. 4). This indicates temperature may have an auto-regressive component to the methane time series. The seasonality

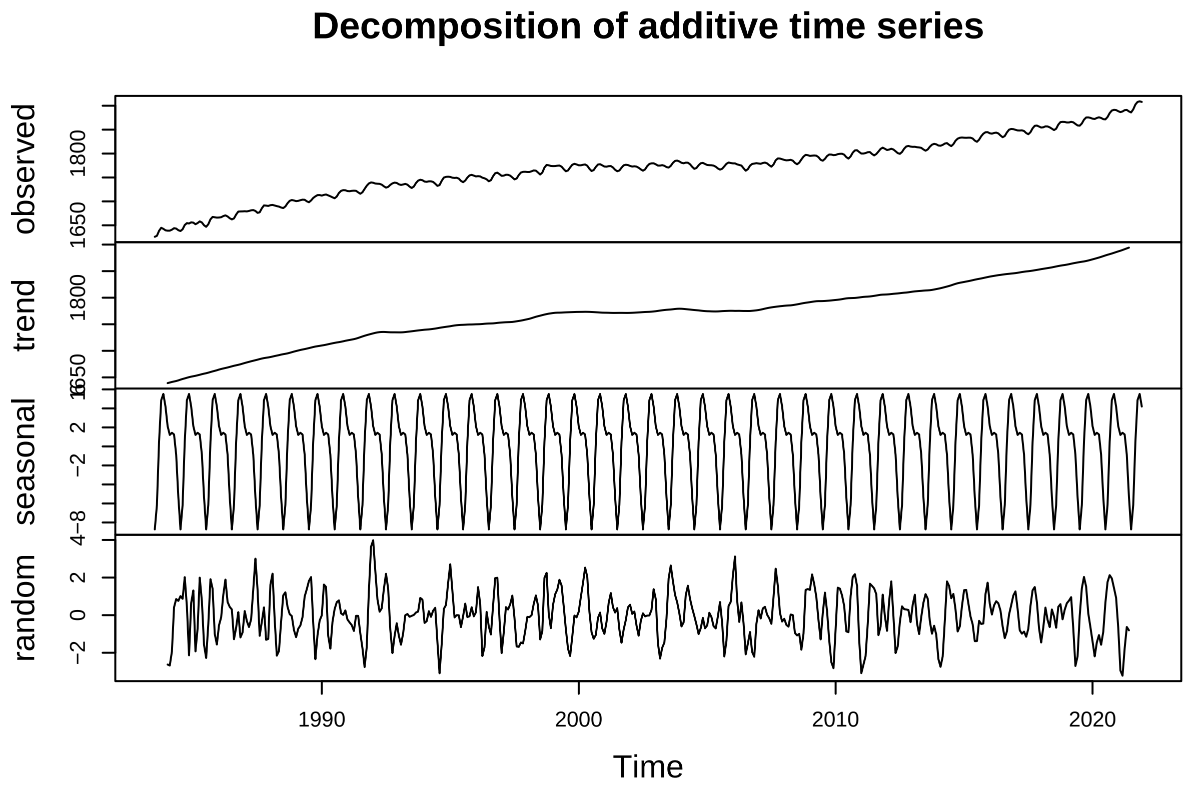


Figure 3: The components of the methane time series

is not strongly visible in the ACF plot.

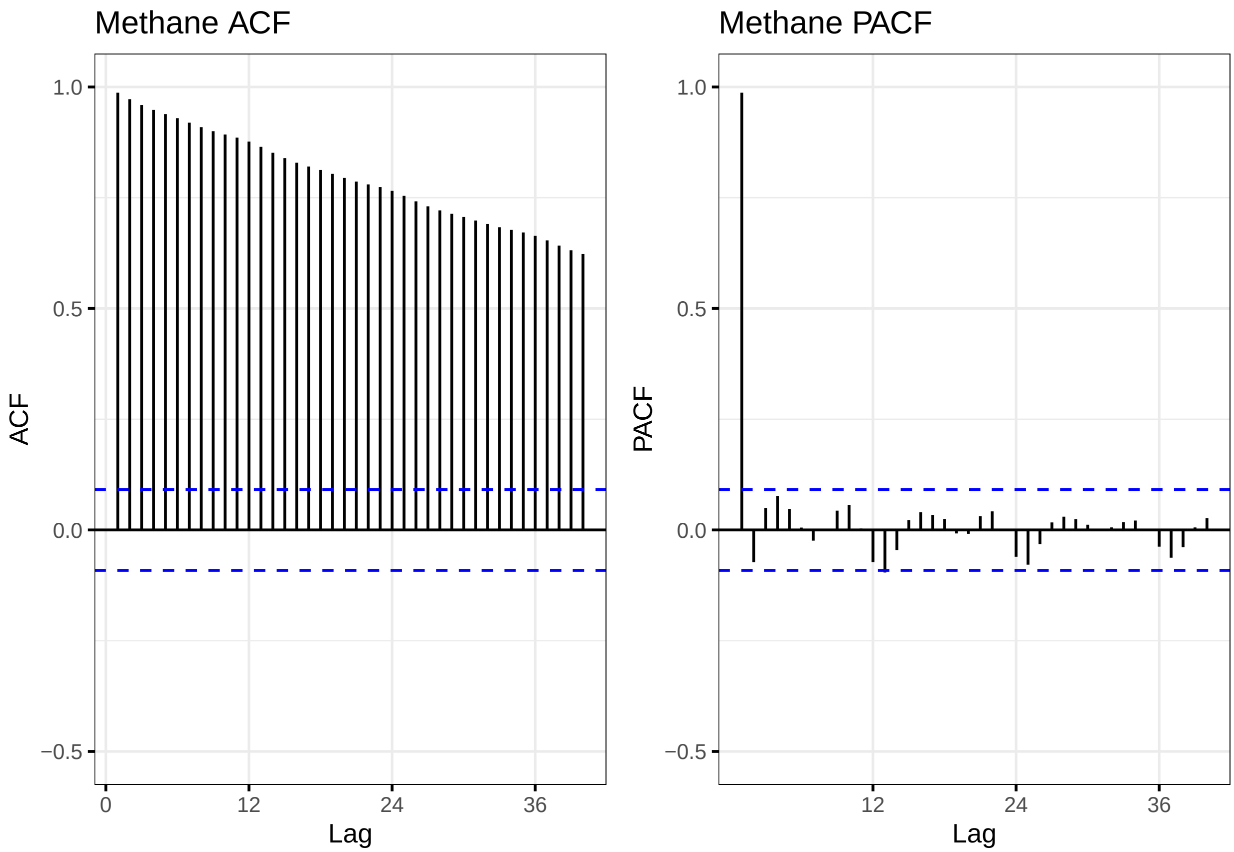


Figure 4: The autocorrelation and partial autocorrelation plots of the methane time series

Before fitting into an ARIMA model, we preformed Augmented Dickey-Fuller (ADF) test and Mann-Kendall test to check for deterministic and stochastic trends in the methane dataset.

```
## [1] "Mann-Kendall Test"
```

```
## Score =  102531 , Var(Score) = 10992237
## denominator =  106491
## tau = 0.963, 2-sided pvalue =< 2.22e-16
```

From the results of the Mann-Kendall test, the p-value is less than 0.05, indicating a significant trend. The test score of 102531 indicates a strong increasing deterministic trend, which agrees with the upward trend in the decomposed series.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  deseasoned
## Dickey-Fuller = -1.7274, Lag order = 7, p-value = 0.6932
## alternative hypothesis: stationary
```

From the ADF test, the p-value of 0.6932 is greater than the significance level of 0.05, meaning that there is not enough evidence to reject the null hypothesis. This series appears to have stochastic trend that changes over time.

Next, the following models were used to predict methane concentrations without the use of exogenous variables.

1. ARIMA using `auto.arima` and adding the decomposed seasonality (from `decompose`) afterwards
2. SARIMA using `auto.arima` with Fourier terms
3. ARIMA with Fourier terms with `auto.arima` function
4. ETS Model with the `stlf` function
5. Neural Network Model with the `nnetar` function
6. TBATS Model with the `tbats` function
7. State Space Model - Exponential Smoothing with the `es` function
8. State Space Model - BSM with the `StructTS` function

The following models were used to predict methane concentrations with the use of exogenous variables (i.e., using the `xreg` term).

As presented in the variable correlation matrix plot, socioeconomic factors exhibit strong colinearity. To deal with this issue, we propose two forms of method:

a) Use a non-linear model (e.g., Neural Network) to incorporate multiple variables simultaneously.
b) Include only one variable at a time in the linear model.

Given the vast number of possible variable combinations, we created a loop function to iterate through all combinations and perform forecasting. We therefore choose the variable which has the lowest `MAPE` score from these loops.

The models with the best predictor is as follows:

1. Arima using `auto.arima` function, using industrial production variable
2. Arima with Fourier terms with `auto.arima` function, using industrial production variable
3. SARIMA using ‘auto.arima with industrial production variable
4. TBATS using `tbats` function, with cpi variable
5. Neural Network Model with the `nnetar` function, with cpi and export variables
6. STL+Arima Model with cpi variable
7. State Space Model with industrial production variable

All of the above models were averaged together in every possible combination. The performance of all of the models were examined with the `accuracy` function and the methane test data subset. All of the best performing models were an average of multiple models both involving and not involving exogenous variables. All of these forecasts closely follow the trend in methane (Fig. 5). **Expand a little more**

Table 4: The top five methane concentration models and their performance statistics sorted by MAPE.

Rank	Model	MAPE
1	ets__nn__nnExoSE__nnFourier__sses__ssesExoSE	0.0540012
2	arimaFourier__nn__nnExoSE__nnFourier__sses__ssesExoSE__stlArimaExoSE	0.0549470
3	arimaFourier__ets__nn__nnExoSE__nnFourier__sarimaExoSE__sses__ssesExoSE	0.0551739
4	ets__nn__nnExoSE__nnFourier__sses__ssesExoSE__stlArimaExoSE	0.0551828
5	arimaFourier__nn__nnExoSE__nnFourier__sses__ssesExoSE	0.0552185

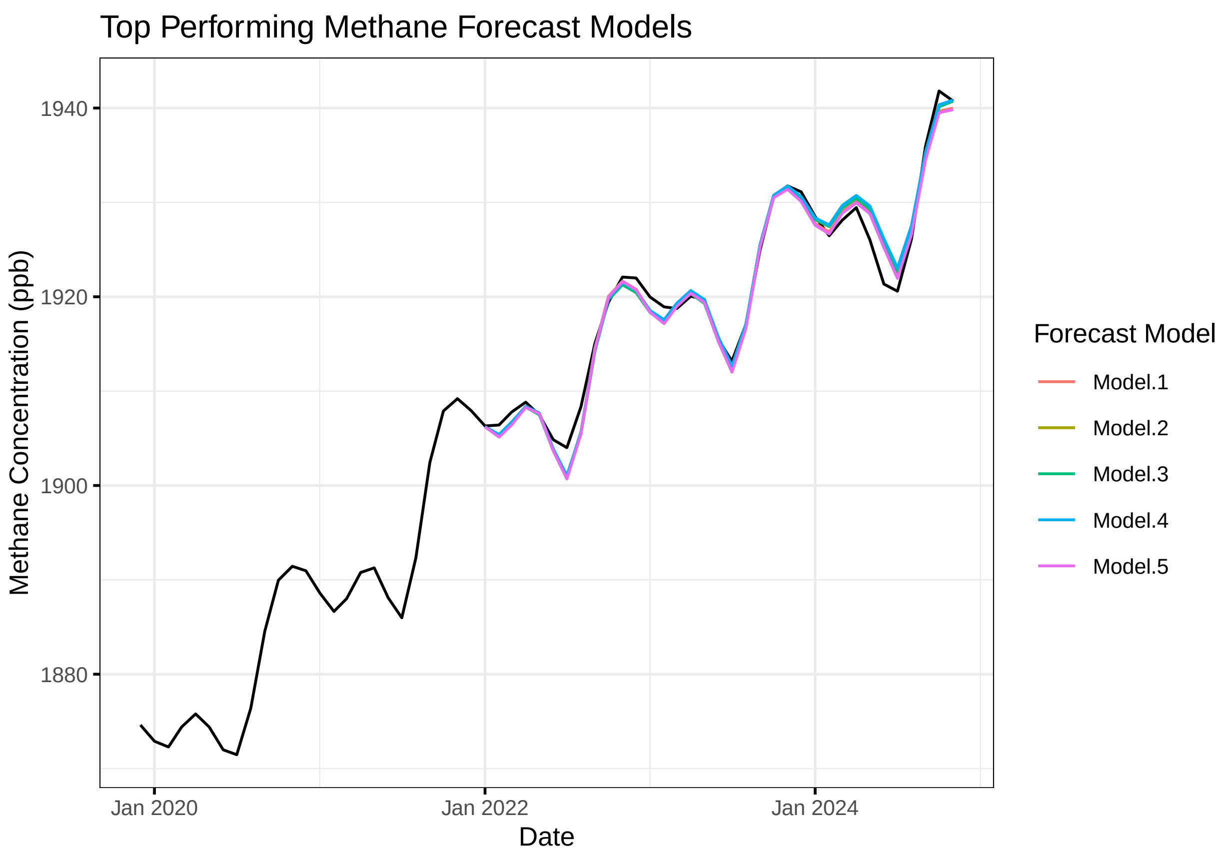


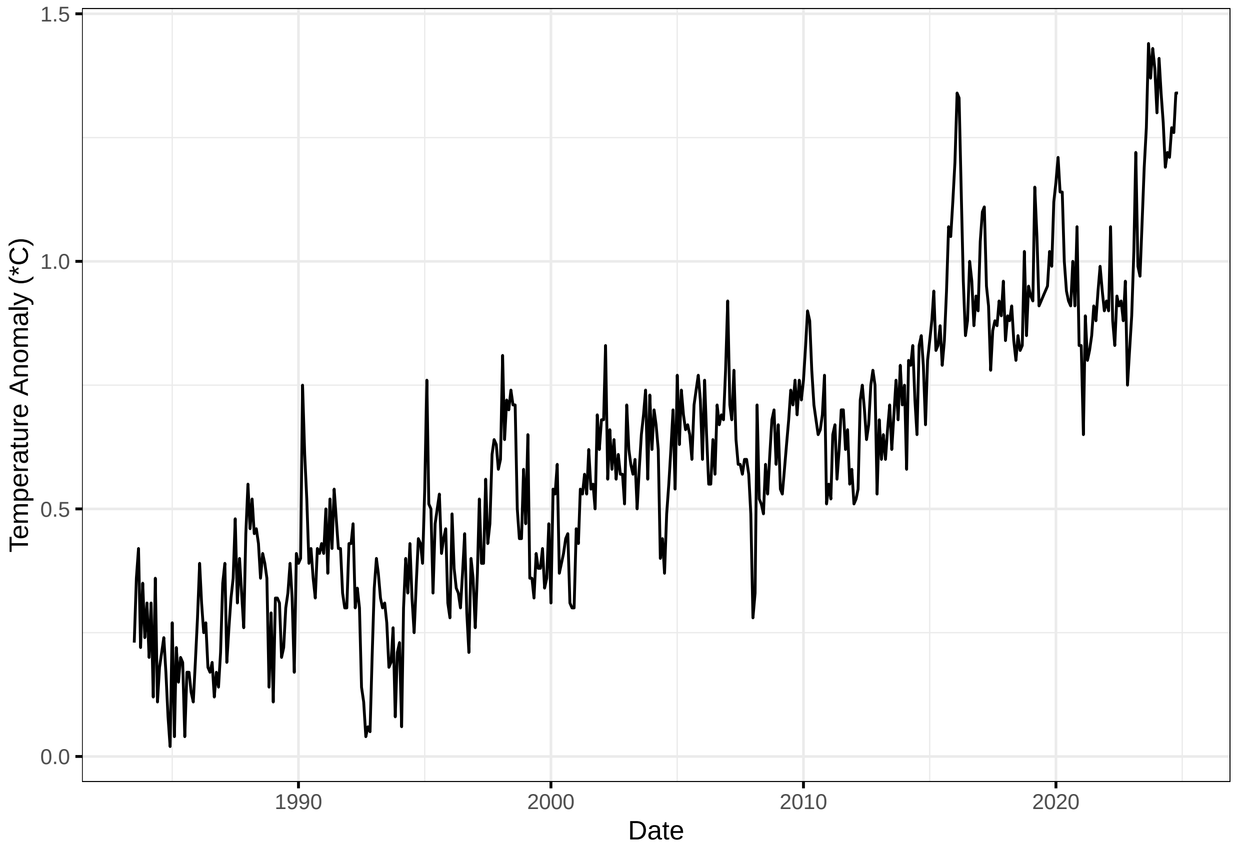
Figure 5: The top five performing methane concentration forecast models. Models are numbered by their performance by MAPE and the full title can be found in Table 4

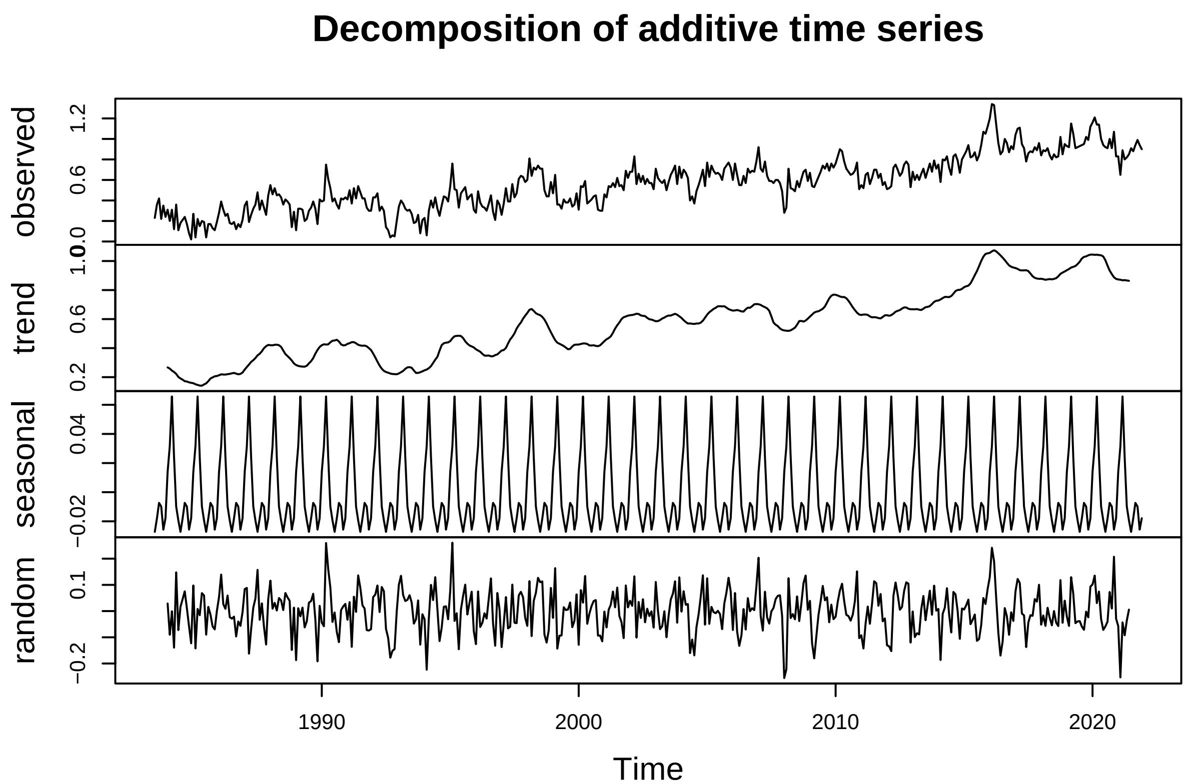
TALK ABOUT TOP 3 MODELS AND LOOK AT THEIR RESIDUALS

IDENTIFY BEST FORECAST

Temperature Anomaly Predictions

Temperature anomaly data was examined using the `decompose` function (Fig. 6 and 7). The temperature anomaly data has a general increasing trend. There is a annual seasonal trend but it is smaller than the random errors and not easily visible. However, it does have a similar shape to the methane seasonality. The remainders do not appear to have any trends and there are few outliers.





The ACF plot shows a gradual decay and the PACF plot shows a sharp drop off after lag 2 (Fig. 8). This indicates temperature may have an auto-regressive component to the model. There does appear to be a slight seasonal scalloping in the ACF plot.

ADD IN STATIONARITY TESTS AND INTERPRET THEIR RESULTS IF TIME

The following models were used to predict methane concentrations without the use of exogenous variables.

1. SARIMA
2. ARIMA with Fourier

The following models were used to predict methane concentrations with the use of exogenous variables (i.e., using the `xreg` term).

1. ARIMAX with Fourier
2. SARIMAX
3. Neural Network
4. Random Forest

All of the above models were averaged together in every possible combination. The performance of all of the models were examined with the `accuracy` function and the methane test data subset. The best performing model was the SARIMAX model. However, the MAPE was much higher than in our methane forecasts (Table 5), and the forecast does not accurately reflect the most recent upwards trends in temperature (Fig. 9). The temperature trended upwards for most of the data record, but recently that trend stabilized. Therefore, the models are likely biased towards recent events and did not capture the shift back to an upwards trend.

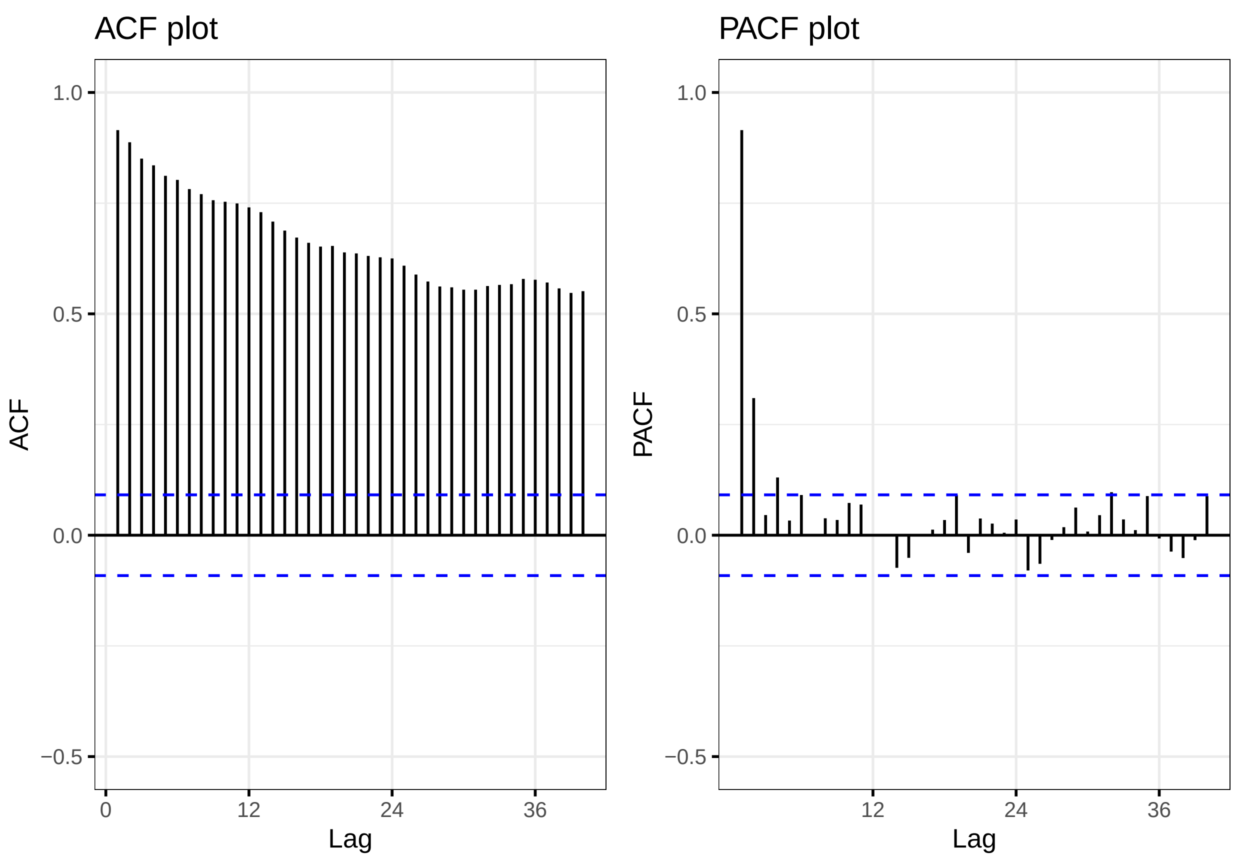


Figure 6: The autocorrelation and partial autocorrelation plots of the temperature time series

Table 5: The top five temperature anomaly models and their performance statistics sorted by MAPE.

Rank	Model	ME	RSME	MAE	MPE	MAPE
1	sarimax	0.0978695	0.2032856	0.1721584	5.932950	14.71139
2	rfExo_sarimax	0.1354008	0.2351881	0.1938629	9.141813	16.11481
3	arimaFourier_sarimax	0.1375237	0.2411710	0.1956617	9.298260	16.14617
4	arimaxFourier_sarimax	0.1375237	0.2411710	0.1956617	9.298260	16.14617
5	sarima_sarimax	0.1510658	0.2437155	0.1979484	10.607708	16.27890

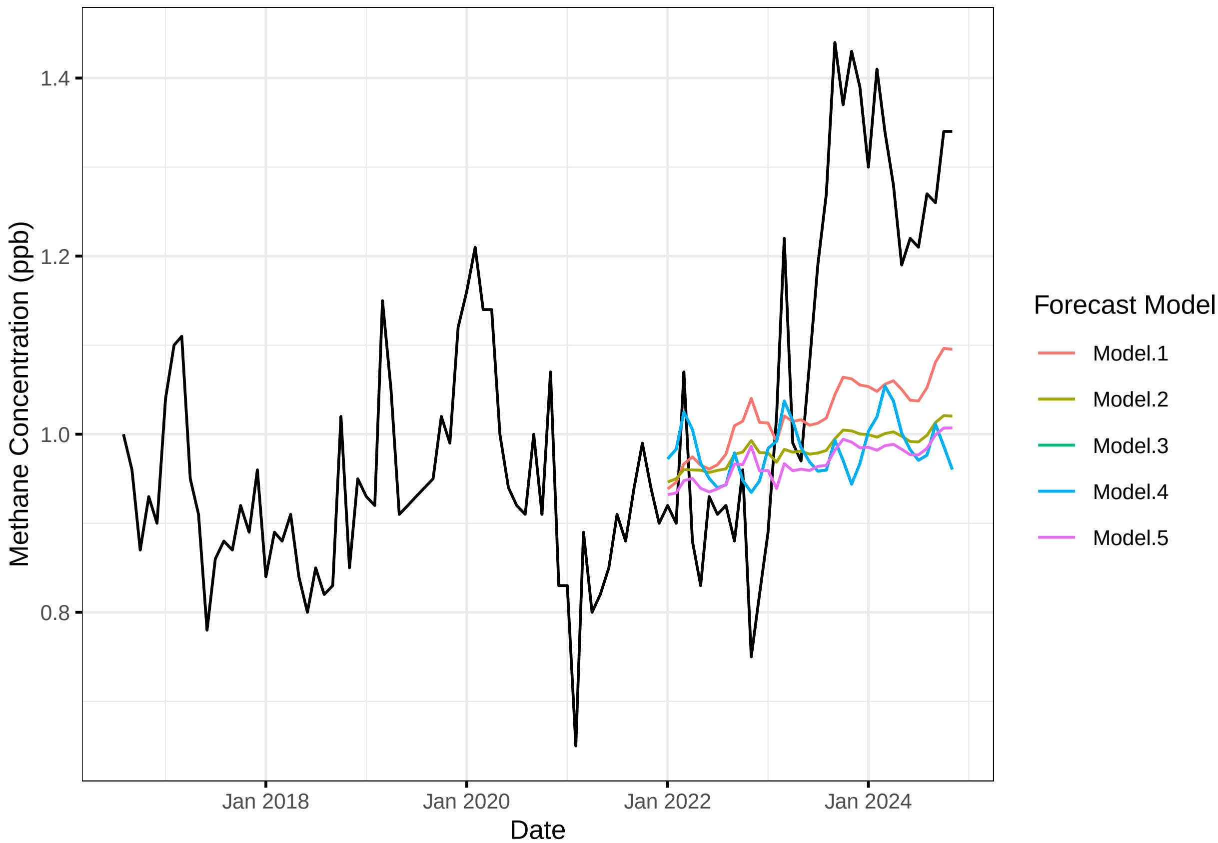


Figure 7: The top five performing methane concentration forecast models. Models are numbered by their performance by MAPE and the full title can be found in Table 5

We used the SARIMAX model to predict temperature for 3 years into the future, using the whole data set as a training data set and assuming that the current climate scenario does not drastically change (Fig. 10). The future project appears to be strongly dominated by the seasonality of methane concentrations with a bias for the stable trend over the past couple of years. It does not seem likely that this forecast will accurately reflect the future temperatures.

Conclusions

Delete this title: Footnotes

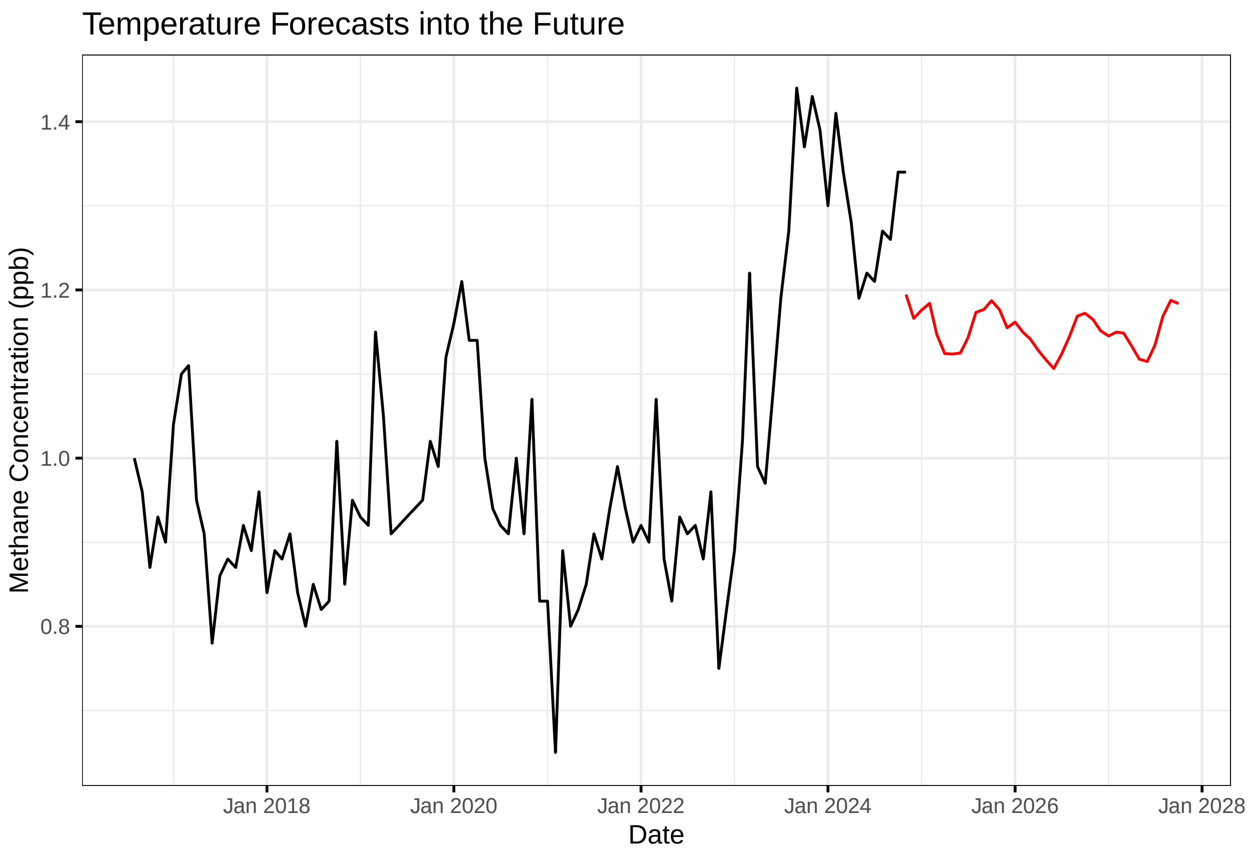


Figure 8: Future temperature forecasts based on the sarimax model with exogenous variable of methane.