POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Empirical model reduction in the study of climate sensibility and variability

Supervisors

Prof. Lamberto RONDONI

Prof. Carlos MEJIA-MONASTERIO

Candidate

Carlo BRUGO

July 2022

Abstract

Predicting climate events means having to deal with a complex, nonlinear, chaotic system, with natural variability in time and space and subject to external forcing. One of the best ways to approach this problem is through the definition of general circulation models, which represent climate systems as a group of subsystems that interact with each other through phenomena. However, they are not always simple to interpret, and sometimes it is necessary to have easy models to read but precise enough to be able to hold all the relevant information.
Empirical model reduction (EMR) is a well-established methodology able to build an efficient model from simple observations of the system. This thesis aims to study the potential of EMR if used to simulate the dynamics of a set of real climate data. After an overview of the basic concepts of climate systems, and an examination of the EMR algorithm, some experiments have been conducted giving the system different properties. The goal is to understand how the algorithm and the data influence the resulting model under variation of some properties of the method, or changing the scenario described by the data. The methodology itself was also compared with alternative versions with more classic data-driven techniques (like linear regression) integrated into the algorithm.
The results obtained help to better understand the conditions under which the empirical model reduction method might be able to substitute a general circulation model with the least possible loss of information.

Acknowledgements

There are some people I need to mention now that my academic path has come to the end.

First of all, those who have eased my path in my master thesis project; prof. Lamberto Rondoni gave me the opportunity to do my project abroad, while prof. Valerio Lucarini and Dr. Manuel Santos Gutiérrez helped me while I was moving my first steps in the climate research field. Above all, I would like to thank prof. Carlos Mejia-Monasterio, who supervised my work during my entire stay in Madrid.

I want to thank my parents for supporting me throughout the ups and downs of my academic path. My sister has then proved to be essential during the difficult times. She was always there for me and we have often shared all the obstacles of university life, supporting each other especially during the lock-down. I must not forget my uncle, who encouraged my interest in technology since I was a little lad. Above all, my biggest university fan: my grandmother. I remember calling her after each exam test, sharing with her all my hopes and fears. Unfortunately, she won't be able to see me graduate, but I'm sure she is smiling from Above, still keeping an eye on me.

I would like to thank all the friends, old and new, which have been part of my life in the last few years. A special mention must go to Davide, Andrea, Gianni, and Chiara, who have often eased the days, even through the worst of times. It could have been much harder without them.

I want to take a moment to mention also some people who have considerably made a difference in my path. Mainly my mentor Francesco, along with the entire community of LeadTheFuture, who taught me to never settle for less than the top when making choices. Meanwhile, my friends of the Silicon Valley Study Tour dream team, led by the tireless Paolo Marenco, helped me discover all the opportunities around us.

The people who know me would tell you that I always have my head in the clouds, so I probably have forgotten to mention someone. I would therefore thank every person who has crossed my path during these years and who has made me the person I am, for better or worse. I dedicate this milestone to you all.

Table of Contents

List of Tables

List of Figures

Acronyms

ACF Autocorrelation Function

ECMWF European Centre for Medium-Range Weather Forecasts

GCM General Circulation Model

GHG Greenhouse Gases

IPCC Intergovernmental Panel on Climate Change

NWP Numerical Weather Predictions

PACF Partial Autocorrelation Function

UNEP United Nations Environment Programme

WMO World Meteorological Organization

Chapter 1

Introduction

1.1 Context overview

Climate science is surely one of the most discussed topics of the last decades. If analyzed from a social, economic, or academic point of view, the climate change problem has surely started to have a deep impact on modern society. In particular, it is possible to affirm that there has been an exponential growth of interest in climate research, starting with the establishment of the Intergovernmental Panel on Climate Change, in charge of the coordination and review of the research activities on the argument.

Among the climate topics currently subject to research, one stands out: general circulation models. The objective is to create models which are capable of realistically simulating the behavior of a climate system, defining it as a group of subsystems that interact with each other. However, there are some issues: due to the numerous subjects and phenomena involved, the climate can be defined as a chaotic system, out of equilibrium and with complex variability. In order to better describe the climate phenomena then, it is necessary to develop a realistic and detailed model, from which it could be really complicated to interpret the results. Another way is to develop a really simple model, but it might be unable to well describe the processes.

To overcome this problem, a new methodology has been introduced: empirical model reduction (EMR) permits to model a physic system and its phenomena as a series of non-linear ordinary differential equations. They describe both the main dynamics of the components as well as their residuals, which might keep some fundamental information. The best part about EMR is that it is a data-driven process, able to parameterize phenomena only using observations.

The recent researches conducted on EMR have all been used as base artificial data, generated using existent physics models described by well-known equations,

such as the Lorenz system. There are few documented cases, if any, of the application of empirical model reduction to real data measures. Since the effectiveness of the method has been proven, it will be interesting to apply it to existent datasets, in order to find possible correlations between the components. There are in fact numerous temporal datasets that provide information about certain climate dynamics. Usually, they are prepared by universities with the aim of using them as input in general circulation models.

1.2 Thesis objectives

The main objectives of this thesis project are the following:

- Provide an overview of the climate research conducted in the last decades, illustrating the core concepts and problems present, as well as the main fields of study.
- Present the empirical model reduction methodology, in particular the algorithm used and its core phases.
- Conduct a series of experiments with the EMR algorithm applied to a set of climate data, varying a series of parameters and commenting on the obtained results.
- Realize different versions of the algorithm using different machine learning techniques, and confront the performances with the original one.

1.3 Main issues

Of course, in the development of the project there are some issues to address:

- There isn't, to date, relevant work about the application of empirical model reduction to real data. There might be some complications and the results might not be as good as the ones obtained in other research.
- Due to the high number of applicable variations in the algorithm, it is impossible to proper describe all of them in this project. Therefore, a simplified version of the algorithm will be used, and only a subset of parameters will be considered.
- An high number of time series could make the system too complex to be well modeled. A small amount of data will be then used, limiting the study to four components.

1.4 Related works

There are some relevant works in the fields of climate, physics, and mathematics which can be studied, in order to better understand the argument.

Talking about climate research, the review realized by Ghil and Lucarini (2020) [1] is a perfect starting point. It reports some of the most important research on the subject, in particular the assessment reports developed by IPCC and some of the more relevant earth system models, like CMIP6 [2]. There have been several works about climate predictions. Some of them exploit the existing GCM and ESM to confirm theories and make predictions [3]. Others search alternative ways, above all machine learning [4], to obtain better predictions [5] and to improve the current models [6, 7].

Speaking about empirical model reduction, it is surely a topic that has grown in popularity in the last years. Several papers have been written on the argument, the majority realized by Ghil, Kondrashov, and Kravtsov [8, 9, 10, 11]. Some of the most recent works have started to confront EMR with other methods, even not data-driven. A great example can be found in the paper realized by Santos Gutiérrez et al. (2021) [12], where EMR is compared to more traditional theory-based approaches, evaluating their performances when used for reduced models parametrization.

Finally, it is considered important to cite some promising new research in the field of dynamical systems: scientific machine learning. The main objective is to exploit machine learning techniques for describing physical phenomena with neural networks, realizing neural differential equations [13, 14] and universal differential equations [15]. Some applications in the field of earth system models are currently in development[1] and soon it will be possible to access promising results.

[1]For further details: `https://github.com/CliMA/ClimaCore.jl`

Chapter 2

Climate Research Overview

Over the years, climate studies have increasingly become a topic of interest, both from an academic and socio-economic point of view. The availability of a growing quantity of observational data and models has led to easier and more detailed access to information. Taking this into account, in 1988 the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP) established the Intergovernmental Panel on Climate Change (IPCC). The IPCC aims to review and coordinate the research activities of the scientific community related to climate change. There are three Working Groups in its structure, each one focalized on a particular topic, which release assessment reports every 4 to 8 years, summarizing the scientific literature relevant to climate change. Independent reviews realized by the research community have also been published over the years, above all the one realized by Ghil and Lucarini (2020) [1], from which will be extracted some basic climate concepts and explained in the next sections[1].

2.1 Basic information on climate science

Overall, climate can be seen as a chaotic system, out of equilibrium and with complex variability. This is due to a series of factors: to begin with, it is composed of a series of phenomena like microphysics of clouds, cloud-radiation interactions, atmospheric and oceanic boundary layers, as well as several scales of turbulence [16]. In addition, its variability is strongly influenced by small forces alteration, both from natural and human-induced sources. It is also subjected to the actions of large-scale agents, which influence the model's evolution. Above all, the absorption

[1]Reprinted figures present in the chapter, with permission from Michael Ghil and Valerio Lucarini (Reviews of Modern Physics, 92.3, p. 035002, 2020). Copyright (2022) by the American Physical Society.

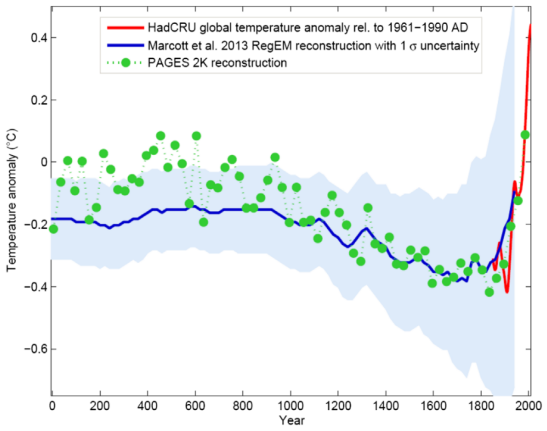


Figure 2.1: Surface air temperature record for the last two millennia. Often called the "hockey stick" it is one of the most discussed graphs in the climate research community. From Ghil and Lucarini (2020) [1].

of solar radiation through the atmosphere leads to the formation of energy fluxes, in order to compensate for the energy disequilibrium formed. These fluxes can be vertical (like infrared radiation directed to the troposphere) in case of a prevalence of absorption at Earth's surface and the atmosphere's lower levels, or horizontal, in case of major absorption on the low latitudes. The vertical and horizontal fluxes form together atmospheric circulation which, coupled with the ocean circulation, reduces the difference between polar and tropics regions, otherwise subject only to solar radiation absorption. Some theories from Lorenz better describe the mechanisms of climate circulation. In particular, Lorenz (1955) [17] illustrates how atmospheric large-scale flows result from the conversion of potential energy (produced by the atmosphere's differential heating) into kinetic energy. Citing the review of Ghil and Lucarini (2020) [1]:
"Overall, the climate system can be seen as a thermal engine capable of transforming radiative heat into mechanical energy with a given, highly suboptimal efficiency given the many irreversible processes that make it less than ideal".

The general dynamic of the climate system phenomenon is usually studied focusing on aspects of different orders of magnitude, but complementary:

- Wavelike features, like planetary waves, which describe phenomena as the transport of energy, momentum, and water vapor.
- Particle-like features, like hurricanes, oceanic vortices, and extratropical cyclones, which affect the local properties of the climate system.
- Turbulent cascades, responsible for the transfer of energy from large to small

scales motion.

Neither of these elements, by themselves, provide a comprehensive understanding of the properties of the climate system.

On top of the difficulties in studying a complex, nonlinear, dynamic system, some additional obstacles are present in this particular case:

- The presence of well-defined subsystems, like the atmosphere, the ocean, and the cryosphere, with distinct physical and chemical properties and different time and space scales.
- Complex processes coupling these subsystems.
- Continuously varying sets of forcings resulted from fluctuations in the solar radiations and natural and human-induced processes.
- Lack of scale separation between processes, which brings difficulties in the use of methods like model reduction and parameterization.
- Lack of detailed, homogeneous, high resolution, and long-lasting observations of climate fields, leading to the reconstruction of missing data using indirect observations.
- The presence of only one realization of the processes of the climate system.

2.2 The data problem

One of the main difficulties in the study of climate systems is surely the lack of homogeneous data, or sometimes lack of data at all. In order to be functional, they need to be of standardized quality, with sufficient temporal coverage and spatially detailed. With this in mind, observation datasets can be divided into two categories: instrumental datasets and proxy datasets.

2.2.1 Instrumental datasets

Instrumental data refer to data obtained by instrumental measurements: first, from the 19th century, with meteorological stations, and then, starting from the 1960s, using remote sensing from satellites (i.e., the Global Observing System of the WMO). Over the years, measurements have grown in number, covering now as much as possible of the entire Earth. However, they are usually sparse, irregular, and of different degrees of accuracy. These problems can be solved with data assimilation approaches: by combining these observations with theoretical

dynamical problems, it is possible to obtain the best estimates of the state.
Three types of problems can be formulated and solved with data assimilation: filtering, smoothing, and prediction. All three of them have the objective of obtaining the best possible estimate of a state $X(t)$ given data in a time interval between t_0 and t_1; with filtering is calculate the estimate of $X(t)$ at time $t = t_1$, with smoothing at $t_0 \leq t \leq t_1$, and with prediction at time $t > t_1$. In Figure 2.2 is possible to see the structure of a forecast assimilation cycle: at determined times $\{t_k : k = 1, 2, \ldots, N\}$, the corresponding state $X(t_k)$ is calculated by combining observations from intervals at $t < t_k$ with the forecast of the previous state $X(t_{k-1})$.
In order to take advantage of the latest improvements in the measurement technology and climate models development, several meteorological centers started to produce reanalyzes, combining archived data with the current best models and data assimilation methods since the 1990s. One of the centers involved in the process of reanalysis, for example, is the European Centre for Medium-Range Weather Forecasts (ECMWF).

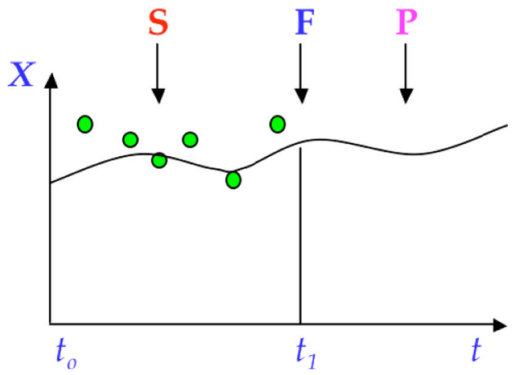


Figure 2.2: Schematic diagram of filtering **F**, smoothing **S**, and prediction **P**; green solid circles are observations. From Ghil and Lucarini (2020) [1].

2.2.2 Proxy datasets

As previously mentioned, instrumental data have a discrete quantity of limitations, above all the fact that direct observations of approximately only the last two centuries are available. It is possible however to indirectly calculate past data about even millions of years ago, thanks to climate proxies. We define the term climate proxies as physical characteristics of the past that have been preserved in various ways and that can be correlated to states of the climate system. Some examples are coral records, tree rings, marine sediment, and ice core. Proxies however vary in terms of precision, uncertainties, and temporal coverage and do

not cover homogeneously the entire Earth. New data assimilation techniques have been then developed, combining simple models with instrumental and proxy data. This has been often addressed as a complex and controversial technique. One major example is the estimation of the globally averaged surface air temperature.

2.3 Climate variability

A perfect way to represent the multiple scales of data in space and time is through the Stommel diagram [18], plotting an idealized spectral density associated with the ocean's variability in logarithmic spatial and temporal scales. In this way, it is possible to describe spatial-temporal variability and associate it with different phenomena, such as cyclones or planetary waves. In Figure 2.3 it is possible to observe that larger spatial scales are associated with larger temporal scales, forming a "diagonal" plot (except for reaching a planet-size spatial scale, from which only the temporal scale increases). Also, it is possible to group scales into three groups: microscale, mesoscale, and synoptic scale. A direct outcome is a complexity in studying models simulating multiple dynamical ranges. Usually, a solution is to examine processes in a particular range, freezing processes with slower timescales. Faster processes, on the other hand, are parameterized. The plot in Figure 2.4 better illustrates climate variability in all timescales, providing semi-quantitative information on the spectral power calculated from several time series. It is possible to observe three types of variability: continuous portions (representing stochastically forced variations), broader peaks (caused by internal modes of variability), and sharp lines (which correspond to periodically forced variations). One of the most particular variability phenomena is El Niño, in the Pacific Ocean: once every 2-7 years, the sea-surface temperature increase by one or more degrees. This particular event is associated with changes in winds and sea-level pressure. Most of the excitement of scientists in climate variability has been caused by variabilities bigger than the interannual one: in particular, paleoclimatic variability (on periods from 10^3 to 10^6 years) can't be fully explained and lots of research are conducted on the topic.

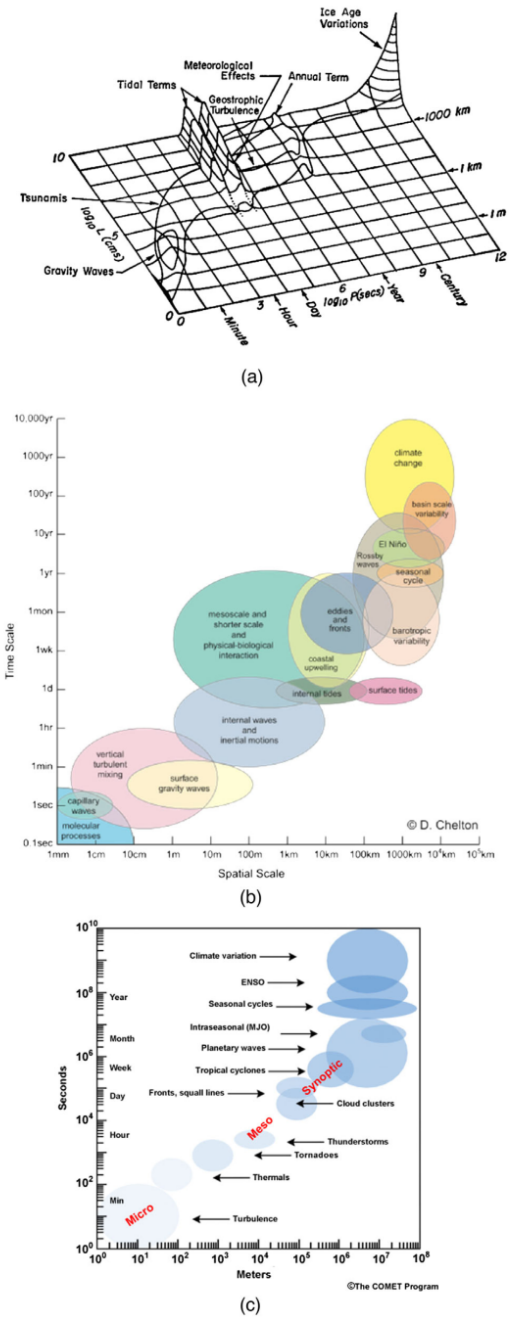


Figure 2.3: Idealized wavelength-and-frequency power spectra for the climate system. (a) The original Stommel diagram representing the spectral density (vertical coordinate) of the ocean's variability as a function of the spatial and temporal scale. (b) Diagram qualitatively representing the main features of ocean variability. (c) The same as (b), describing here the variability of the atmosphere. From Ghil and Lucarini (2020) [1].

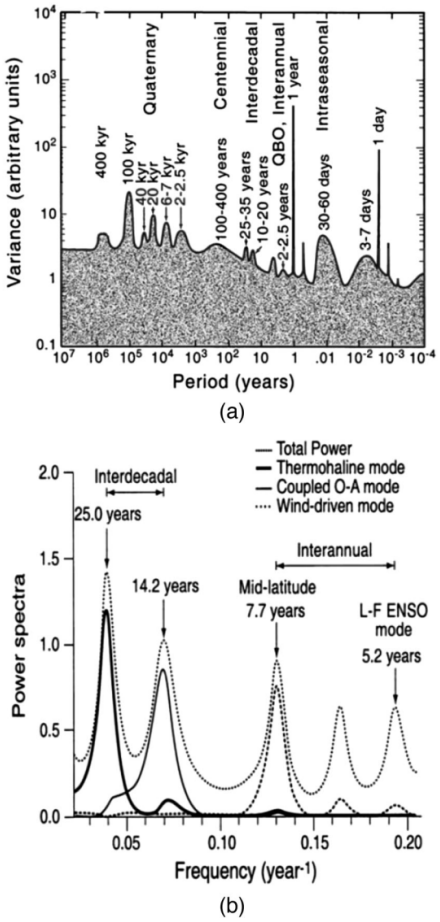


Figure 2.4: Power spectra of climate variability across timescales. (a) An artist's rendering of the composite power spectrum of climate variability for a generic climatic variable, from hours to millions of years; it shows the amount of variance in each frequency range. (b) Spectrum of the Central England temperature time series from 1650 to the present. Each peak in the spectrum is tentatively attributed to a physical mechanism. From Ghil and Lucarini (2020) [1].

2.4 Climate predictions and climate models

A key area in climate science (and focus on this thesis) is the development of general circulation models (GCM), which simulate the behavior of a climate system. In particular, they try to simulate in the best possible way all the climate subsystems (atmosphere, ocean, and land) and their interactions. They are the base for Earth system models, which also integrate the biological subsystem. In Figure 2.5 is possible to examine all the components of an Earth System Model, along with the

evolution of climate models across the first four IPCC.

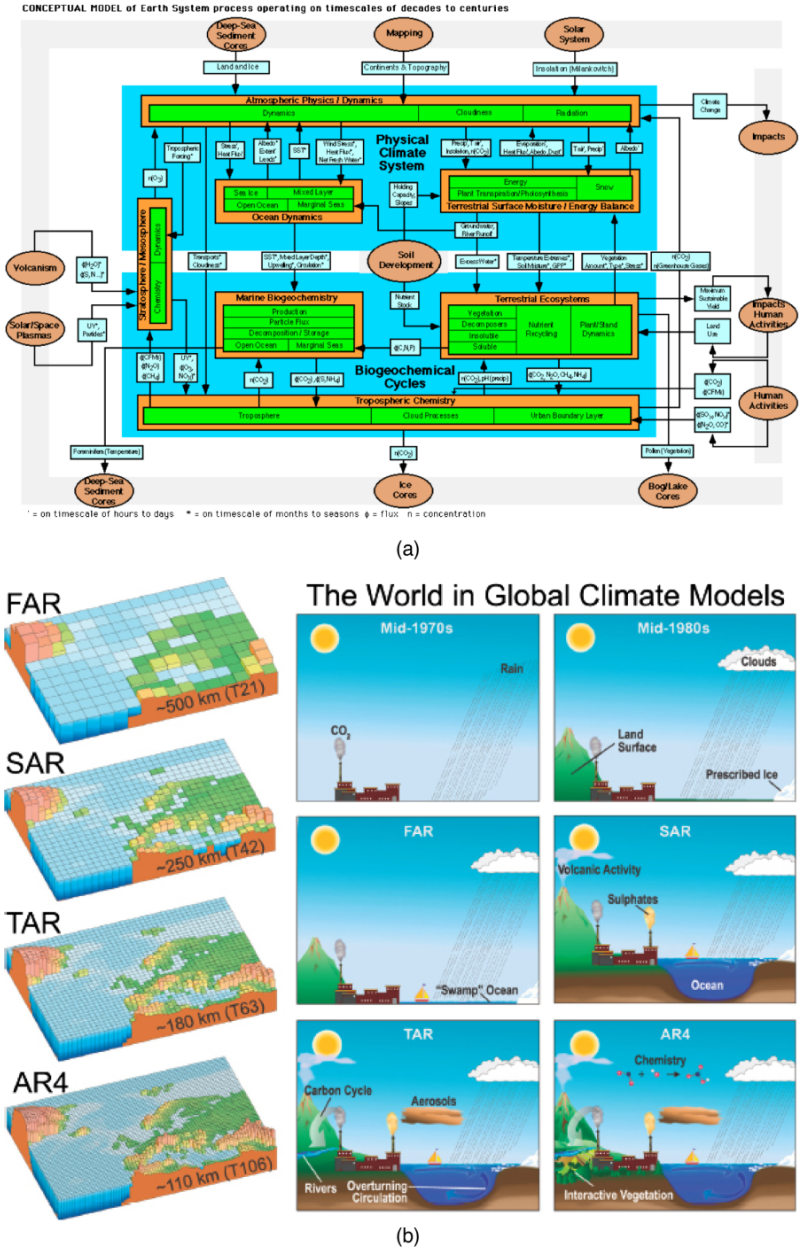


Figure 2.5: The Earth system, its components, and its modeling. (a) The NAC (1986) horrendogram that illustrates the main components of the Earth system and the interactions among them. (b) Evolution of climate models across the first four IPCC assessment reports, ranging from the early 1990s to the mid-2000s. From Ghil and Lucarini (2020) [1].

One of the main problems with using climate modeling to make predictions is that simulating chaotic events is subjected to uncertainties of the first kind (as defined by Lorenz in 1976 [19]): a small error in the initial data can lead to bigger uncertainties in the prediction. Managing and reducing these uncertainties are some of the goals in the field of numerical weather predictions (NWPs). One of the main solutions found right now is the use of slightly perturbed initial states, which permits to study a probabilistic estimate of the system evolution.
Another problem arises in the presence of uncertainties in the model formulation, called uncertainties of second kind. They are related to the presence of uncertainties in key parameters of the climate system (called parametric uncertainties), as well as a possible poor representation of certain properties (causing structural uncertainties).
Over the years, more and more climate models have been developed; one of the IPCC reports of 2014 lists about 50 models. Most of these models, however, share a fair quantity of components, since the main part of them originated from a small number of atmospheric and oceanic models, and this brings to also similarities in their behaviors and results. To better manage these resemblances and to improve comparisons among distinct models, the Program for Climate Model Diagnostics and Intercomparison (PCMDI), defines a series of standards for the modeling research through its Climate Model Intercomparison Projects (CMIPs). A typical series of initial data used for evaluating a climate model for IPCC reports typically includes:

- A reference state, like a preindustrial state with fixed parameters.
- Industrial era and present-day conditions, including natural and anthropogenic forcings.
- Future climate scenarios, using a set of future scenarios of greenhouse gas, aerosol emissions, and land-use change.

Any new variation of data, like greenhouse gases, brings the system to a new stationary state. An evolution from one state to another is called a scenario. Each new scenario is a representation of the expected greenhouse gas and aerosol concentrations, resulting from an industrialization path and change in land use.
The CMIP permitted to bring also standardization of metrics to statistically estimate the model performance. The validation of a model can be divided into two distinct operations:

- Model intercomparison, to assess the consistency of different models in the simulation of certain physical phenomena in a certain time frame.
- Model verification, comparing a model's output with the corresponding observed or scalar quantities.

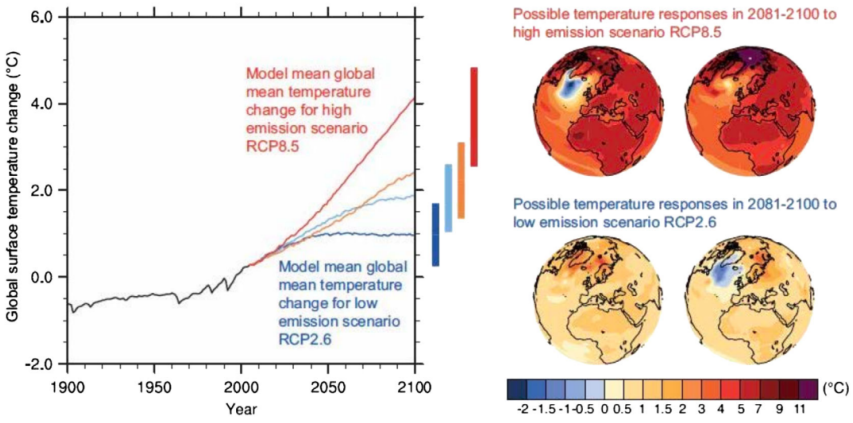


Figure 2.6: State-of-the-art climate model outputs for various climate change scenarios. (Left) Change in the globally averaged surface temperature as simulated by climate models included in the IPCC assessment report 2014a. Vertical bands indicate the range of model outputs. (Right) Spatial patterns of temperature change, i.e., a 2081–2100 average with respect to the present. From Ghil and Lucarini (2020) [1].

Choosing the most suitable metric means constructing any suitable function from the model's variables. The problem is that variables can have physical relevance and robustness which can differ widely, and it is impossible to validate them a priori. So, different metrics are necessary in different cases. Some other issues which can create problems in the model validation are:

- The presence of three types of attractors: an attractor of the real climate system, its reconstruction from observations, and the attractors from the climate model.
- The high dimensionality of the phase space and parameter space of the attractors.

To address these issues, multivariate metrics and multi-model simulations have been used, in order to have a better overview of the climate system.

2.5 Climate sensitivity and response theory

One of the main objectives of climate studies has always been predicting the impact of changes in the climate system's parameters, like the variation of greenhouse gas concentration. In particular, it's indicated as climate sensitivity the evaluation of the response of the climate system to external perturbations. One of the main

applications is the projection of temperature changes over the coming years caused by the increasing concentration of greenhouse gas.
To make a simple example, it's possible to consider an energy balance model equation, describing the evolution of surface air temperature. The simplest zero-dimensional model (with "zero" indicating the number of independent space variables used to describe the model domain) can represent the evolution of temperature with the following equation:

$$c\frac{dT}{dt} = R(T) = R_i(T) - R_o(T)$$

with T being the surface temperature, c the global atmospheric and oceanic heat capacity, and R the net radiation, calculated as the difference between the incoming solar radiation R_i and the outgoing terrestrial radiation R_o. In this context, we can consider the difference in global annual mean temperature ΔT between two statistical steady states, with distinct CO_2 concentration levels. Assuming that changing the CO_2 concentration corresponds to applying an extra net radiative forcing $\Delta\tilde{R}$ to the system, it will be a corresponding change ΔT so that $R(T_0 + \Delta T) + \Delta\tilde{R} = 0$.
Several aspects can affect net radiation, like cloud cover and greenhouse gas concentration. Above all, the main reference factor is the CO_2 concentration. In fact, it is defined as equilibrium climate sensitivity (ECS) the globally and annually averaged increase of surface temperature due to the doubling of the concentration of CO_2 with respect to a reference state. Several studies on ECS have been conducted in the last years, exploiting mainly tools like Earth System Models. Most of this scientific research has its base on the Ruelle response theory and its derivatives.

2.5.1 Linear response theory and applications on climate systems

The Ruelle response theory [20] contributed a lot to address problems, like the study of climate sensitivity, in the setting of dynamical systems theory, rather than in statistical mechanics. It makes it possible to compute changes in a particular measure x due to weak perturbations of intensity ϵ. We can then write:

$$\dot{x} = F(x,t) = F(x) + \epsilon X(x,t)$$

where $F(x)$ is the background dynamics of x and $\epsilon X(x,t)$ its perturbation.
We can then evaluate the expectation value $\langle\Psi\rangle^\epsilon(t)$ of a measurable observable $\Psi(x)$ with the formula:

$$\langle\Psi\rangle^\epsilon(t) = \langle\Psi\rangle_0 + \sum_{j=1}^{\infty} \epsilon^j \langle\Psi\rangle_0^{(j)}(t)$$

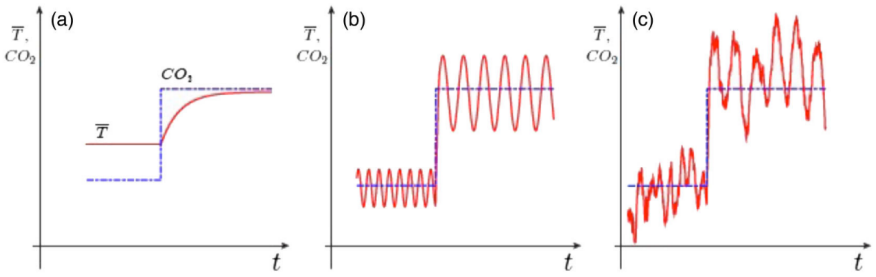


Figure 2.7: Climate sensitivity (a) for an equilibrium model, (b) for a nonequilibrium oscillatory model, and (c) for a nonequilibrium model featuring chaotic dynamics and stochastic perturbations. As a forcing (atmospheric CO_2 concentration, blue dash-dotted line) changes suddenly, global temperature (red thick solid line) undergoes a transition. (a) Only the mean temperature $\overline{T}$ changes. (b) The amplitude, frequency, and phase of the oscillation change too. (c) All details of the invariant measure, as well as the correlations at all orders, are affected. From Ghil and Lucarini (2020) [1].

In this context, we talk about linear response theory if we apply only a linear perturbation, with the case $j = 1$, so:

$$\langle \Psi \rangle^{\epsilon}(t) = \langle \Psi \rangle_0 + \epsilon \langle \Psi \rangle_0^{(1)}(t)$$

This is a common simplification of the theory due to the fact that, with $\epsilon < 1$, the component of the formula $\epsilon^j \langle \Psi \rangle_0^{(j)}(t)$ will converge to 0 with $j \to \infty$.
Many are the scientific research conducted in the recent years regarding the verification of linear response theory, one above all realized by Lucarini, Ragone, and Lunkeit in 2017 [3]. One of the possible future subjects might also be the application of exact response theory [21] to the study of climate.

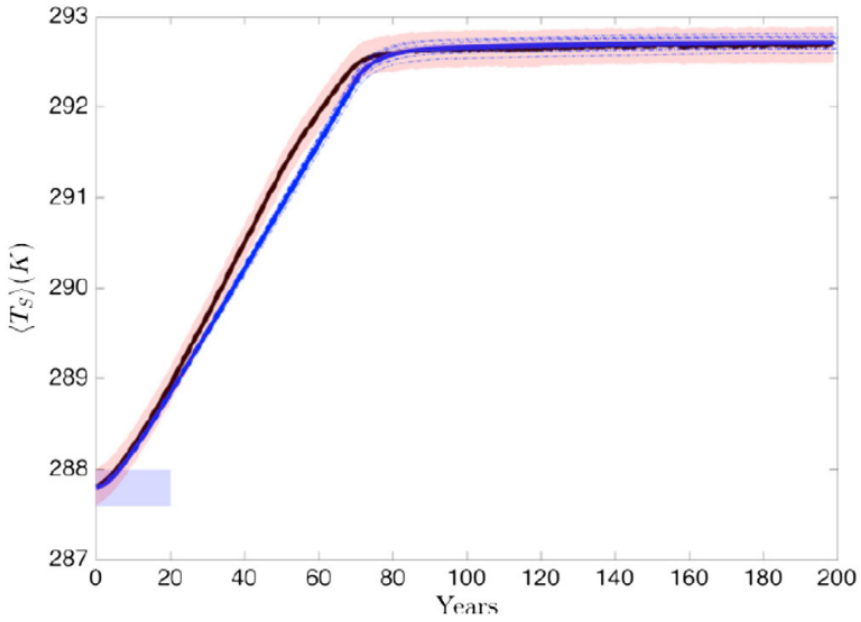


Figure 2.8: Comparison between the climate model simulation (black line) and response theory prediction (blue line) for a experiment using a PlaSim model. The CO_2 concentration was ramped up by 1% per year to double its initial value. From Ghil and Lucarini (2020) [1].

Chapter 3

Empirical Model Reduction

One of the first assumptions discussed in Chapter 2 is the fact that climate systems are complex, non-linear, and dynamic. In order to better simulate the processes involved, complex general circulation models are built. However, a few problems must be taken into account; the first one is in the interpretation of the results: the evolution of climate phenomena developed in a detailed and realistic model is more difficult to understand. On the other hand, highly simplified models can help to better understand isolated processes, but not their interaction. The second problem is that some unresolved processes involving the dynamical variables of interest might be difficult to be parameterized and, consequently, lost in the simulation. It's important to consider that these observations can be done on any dynamical system involving stochastic processes. A solution for both problems is the realization of a model of intermediate complexity, resolving a subset of climate systems and parameterizing the unresolved part as stochastic processes. Empirical model reduction (EMR) [8, 10, 11] is a methodology perfect for the scope. It is also simple to implement since it is able to construct models based almost entirely on observational data of the system, both in the case of actual observational datasets or results from high-end simulations. The EMR methodology is typically used to build multilayer stochastic models, which try to represent the dynamics of a group of observed and unobserved variables using only available data. An example in using empirical model reduction can be found in Appendix A, where is applied to a Lorenz system.

3.1 Definition

A better technical definition of the construction of a EMR model will be shown in this section. Let's consider a multivariate time series $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n})$ with n the number of its components. Indicating with $\overline{\boldsymbol{X}}$ its time mean and $\boldsymbol{x} = \boldsymbol{X} - \overline{\boldsymbol{X}}$

the vector of its anomalies, it is possible to define the evolution of $\boldsymbol{x}$ as:

$$\dot{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{x} + \boldsymbol{N}(\boldsymbol{x})$$

where $\boldsymbol{L}$ is a linear operator and $\boldsymbol{N}$ represents the nonlinear components. With this in mind, it's important to consider two aspects: (1) the formula above is really complicated to resolve, even having its exact form and (2) $\boldsymbol{x}$ is always a sum of the ideal signal $\boldsymbol{x_S}$ and a noise component $\boldsymbol{x_N}$. One of the most efficient ways to address these problems is through a data-driven approach, but it is necessary to be careful. As said at the beginning of the chapter, there could be some unobserved components that contribute to the dynamics of the observed ones and it is important to take them also into account. With the EMR methodology, it is possible to easily build an inverse multi-level stochastic model, which relies almost entirely on the observations, while making assumptions about the underlying dynamic, described as residual components $\boldsymbol{r}$. The new model results in a set of ordinary differential equations, the first one describing the dynamics of the observable variables and the others its residual components.

$$\begin{aligned} d\boldsymbol{x} &= \boldsymbol{C}dt - \boldsymbol{L}\boldsymbol{x}dt + \boldsymbol{N}(\boldsymbol{x},\boldsymbol{x})dt + \boldsymbol{r}_1 dt, \\ d\boldsymbol{r}_l &= \boldsymbol{C}_l dt + \boldsymbol{M}_l(\boldsymbol{x},\boldsymbol{r}_1,\ldots,\boldsymbol{r}_l)dt + \boldsymbol{r}_{l+1}dt,\ 1 \le l \le L-1, \\ \boldsymbol{r}_L dt &\approx \Sigma dW \end{aligned} \tag{EMR}$$

In the equation above the vertical vectors $\boldsymbol{C}$ and $\boldsymbol{C}_l$, and the matrices $\boldsymbol{L}$, $\boldsymbol{N}$, and $\boldsymbol{M_l}$ represent the coefficients to compute from the observational data. $d\boldsymbol{x}$ and $d\boldsymbol{r}_l$ indicate the tendencies of the time series $\boldsymbol{x}$ and $\boldsymbol{r}_l$ respectively. They are calculated, component-wise, with the following formula:

$$\begin{aligned} dx(t) &= x(t+dt) - x(t), \\ dr_l(t) &= r_l(t+dt) - r_l(t),\ 1 \le l \le L-1 \end{aligned}$$

As said before, the dynamics related to the unobserved variables, as well as the noise components, are acknowledged by adding the regression residuals described with an additional level. Ideally, levels are added until the last residual $\boldsymbol{r}_L$ can be reasonably approximated as spatially correlated white noise with a spatial covariance matrix Σ. As explained in Appendix A of [11], this can be established by analyzing the currently last level residual. If well approximated by a white noise, it should de-correlate at lag dt. Performing the regression of $d\boldsymbol{r}_{L-1}$, all the coefficients should then approach 0, except the one corresponding to $\boldsymbol{r}_{L-1}$ which should be -1. The result is:

$$d\boldsymbol{r}_{L-1}(t) = \boldsymbol{r}_{L-1}(t+dt) - \boldsymbol{r}_{L-1}(t) \simeq -\boldsymbol{r}_{L-1}(t) + \boldsymbol{r}_L(t)$$

The new residual $\boldsymbol{r}_L$ is then a lagged copy of $\boldsymbol{r}_{L-1}$. The direct consequence is that the coefficient of determination R_i^2 of the i-th component of the time series at level $L-1$ becomes:

$$\begin{aligned} R_i^2 &= 1 - \frac{\sum_k r_{L,i}(k)}{\sum_k (r_{L-1,i}(k+1) - r_{L-1,i}(k))^2} \\ &\simeq 1 - \frac{\sum_k r_{L-1,i}(k+1)}{\sum_k (r_{L-1,i}^2(k+1) + r_{L-1,i}^2(k))} \simeq 1 - \frac{var(r_{L-1,i})}{2var(r_{L-1,i})} = 0.5 \end{aligned}$$

In the end, the best way to determine the best number of levels L is by verifying when the coefficient of determination converges with all the components of the time series.

3.2 The empirical model reduction algorithm

In this section will be reported a description of the main steps of the algorithm used to build an empirical model reduction model given a multivariate time series. First of all, it is important to notify that the algorithm used in this project is a simplified version of the original one. Its first version is available as a MATLAB package on the website of the Department of Atmospheric and Oceanic Sciences of the University of California, Los Angeles[1].

3.2.1 Differences from the original version

The main difference between the original version and the new one is that the code has been translated entirely to another programming language, from MATLAB to Python 3. This choice has been made for two main reasons: (1) to exploit powerful Python libraries like `numpy`, which permits performance optimizations like vectorization and multi-threading operations, and (2) to better manage operations and the analysis of results using Jupyter Notebooks (further details in Chapter 4). In its original form, the empirical model reduction was executed in a MATLAB function, where the coefficients were calculated from the dataset and a new multivariate time series was constructed from the model.

During the translation, some changes in the algorithm design have been made, in order to make it simpler for the case of study:

- The possibility of adding an external periodic forcing to the time series has been removed.

[1]Package available at the following url: `http://research.atmos.ucla.edu/tcd/dkondras/Software.html`

- The original algorithm gave the possibility to use another method, instead of the matrix linear equation solving, to calculate the coefficients of the equations (in this case partial least squares regression). This approach has been generalized in the new algorithm, giving the possibility to define an alternative function for calculating the coefficients and passing it as an argument to the EMR one.
- The residual of the last level $\boldsymbol{r}_L$ is always considered as spatially correlated white noise, following the assumptions of the previous section. In the original code, other possibilities (like the absence of $\boldsymbol{r}_L$) are considered.
- All the components of the reconstructed simulated data are returned, while in the original code there was the option to select a subset of them.
- The possibility to execute the empirical model reduction to a sub-sample of the time series has been added inside the function, while before it was necessary to perform the action before calling the function.

In addition to these modifications, the code has received some adjustments, in order to make it more performing and easier to understand.

3.2.2 Algorithm description

An overview of the EMR function can be seen on Algorithm 1. In this section all the step swill be discussed, dividing them into three parts: the initial setup, the model construction, and the simulation of the new time series from the model. It's important to notify that the program shown is a simplified version, able to build equations only up to the second degree, while the one used for the tests is able also to build equations of third and fourth degree.

Input and output

There are six input data given to the function in order to construct an EMR model:

- ***data***, a $m \times n$ matrix containing n time series from which construct the new model.
- *n_level*, which indicates the number of levels L (main plus residuals) of the model.
- *eq_deg*, indicating the degree of the equation of the main level.
- K, which is the coefficient of sub-sampling; depending on its value, one every K observations are taken into account.

- n_iter, indicating the number of successful simulations to run.
- len_{sim}, the length of the simulated time series (typically equivalent to $\frac{m}{K}$, but can be longer or shorter as necessary.

The main output data returned from the function are a matrix sim_data, containing a multivariate time series for each successful simulation, and a matrix with the coefficients of determination R^2 of all the components.

Initial setup

First of all, the initial time series is sampled, selecting one observation every K. Then, a matrix x is initialized with the values of the time series. It will be updated during the program execution, containing at the end, for each time t, the values of the time series of each component and the corresponding residuals r, with the format:

$$\boldsymbol{x} = \begin{bmatrix} x_1 & \dots & x_n \\ r_{1,1} & \dots & r_{1,n} \\ \dots & \dots & \dots \\ r_{L-1,1} & \dots & r_{L-1,n} \end{bmatrix}$$

Two variables are then declared: $\boldsymbol{A}$ and $\boldsymbol{x_A}$, which will contain respectively all the coefficients to calculate and their corresponding multipliers. In fact, in order to construct the differential equations of (EMR) in a more efficient way, all the coefficients are collected in a unique matrix. The idea is to exploit vector calculus building the equations in the form:

$$\boldsymbol{f_0} = \boldsymbol{C} - \boldsymbol{L}\boldsymbol{x} + \boldsymbol{N}(\boldsymbol{x}, \boldsymbol{x}) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j \geq i} a_{i,j} x_i x_j,$$

$$\boldsymbol{f_l} = \boldsymbol{C}_l + \boldsymbol{M}_l(\boldsymbol{x}, \boldsymbol{r}_1, \dots, \boldsymbol{r}_l) = a_{rl,0} + \sum_{i=1}^{n} a_{rl,i} x_i + \sum_{j=1}^{l} \sum_{i=1}^{n} a_{rl,(j,i)} r_{j,i},\ 1 \leq l \leq L-1$$

and simplifying them with a matrix multiplication for every level l

$$\boldsymbol{f_l} = \boldsymbol{A}[l] \times \boldsymbol{x_A}[l],\ 0 \leq l \leq L-1$$

In this way, it is also easy to calculate the residuals with the formulas

$$\begin{aligned} \boldsymbol{r}_1 &= d\boldsymbol{x} - \boldsymbol{f_0}, \\ \boldsymbol{r}_{l+1} &= d\boldsymbol{r}_l - \boldsymbol{f}_l,\ 1 \leq l \leq L-1 \end{aligned} \tag{r}$$

The shape of $\boldsymbol{A}$ is determined by the maximum number of coefficents for each level, which is the higher value between $\prod_{i=1}^{n} \frac{eq\_deg+i}{i}$ (number of coefficients on the main level equation) and $n\_level * n + 1$ (number of coefficents on the last residual equation).

EMR model construction

The model construction is performed starting from the first level to the subsequent ones. First of all, the data of the current level are normalized on each component using a standard score, by subtracting their mean and dividing them by their standard deviation. Then their discrete derivative is computed with the formula

$$\boldsymbol{dx}[t] = \boldsymbol{x}[t+1, l] - \boldsymbol{x}[t, l]$$

In order to exploit the vectorized form of (EMR), it is necessary to construct the matrix $\boldsymbol{x_A}$ with the required values, so

$$\begin{aligned}
&\boldsymbol{x}_A[l] = [x_1, \ldots, x_n], && main\,level,\ eq_deg = 1\\
&\boldsymbol{x}_A[l] = [x_1, \ldots, x_n, x_1^2, \ldots, x_1 x_n, \ldots x_n^2], && main\,level,\ eq_deg = 2\\
&\boldsymbol{x}_A[l] = [x_1, \ldots, x_n, r_{1,1}, \ldots, r_{1,n}, \ldots, r_{l-1,n}], && residual\,level
\end{aligned}$$

We can then calculate the coefficients in $\boldsymbol{A}[l]$ resolving the matrix multiplication analyzed in the initial setup section, with f_l equal to $d\boldsymbol{x}$ in the main level and $d\boldsymbol{r}_l$ in the residual ones. In order to resolve the equation, it is possible to calculate the inverse of $\boldsymbol{x_A}[l]$ using the singular value decomposition factorization [22], which decompose the original matrix into three matrices $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$. $\boldsymbol{S}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{x_A}[l]$ and $\boldsymbol{U}$ and $\boldsymbol{V}$ are two orthogonal matrices. On the main level, for example, the coefficients are calculated with the following equation:

$$\begin{aligned}
&\boldsymbol{dx'} = \boldsymbol{dx} - \mu_{dx}, \quad \boldsymbol{x'}_A = \boldsymbol{x}_A[l] - \mu_{x_A,l},\\
&\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}^T = SVD(\boldsymbol{x'}_A),\\
&\boldsymbol{A}[l] = (\boldsymbol{x'}_A)^{-1} \times \boldsymbol{dx'} = \boldsymbol{V} \times \boldsymbol{S}^{-1} \times \boldsymbol{U^T} \times \boldsymbol{dx'}
\end{aligned}$$

The parameter a_0 is then separately computed with the following formula and added to $\boldsymbol{A}[l]$.

$$a_0 = \mu_{dx} - \boldsymbol{A}[l] \times \mu_{x_A,l}$$

Once calculated the coefficients, it is then possible to calculate the residuals with the formula (r) and storing them into the matrix $\boldsymbol{x}$ for the next level iteration. Finally, the R^2 coefficients for each component are calculated from the residuals r and the discrete derivative dx.

Model simulation

Once calculated the EMR coefficents, it is possible to use them to compute a simulation of the data, which can be compared to the original one in order to study the model performance. The initial data at time $t = 1$ are taken from the original one and saved in a matrix $\boldsymbol{x}_{step}$. After that, the vector $\boldsymbol{x}_A$ is calculated in the same way done in the previous section of the algorithm, but adding an initial 1 which will correspond to the coefficient a_0 in the matrix multiplication. After the first step, an iteration cycle starts. First of all, the data in the current step are exported to ***sim_data*** but, if some value of $\boldsymbol{x}_{step}$ diverged, the simulation is considered failed and a new one is initialized. Then, the value of the new time step is calculated. Starting with the residuals, almost all of them are obtained from $\boldsymbol{x}_{step}$, while the last one is considered, like said previously in the chapter, as a correlated white. This is possible calculating the Cholesky matrix[2] of the residuals on the last level calculated during the model construction and applying on them the Monte Carlo method. Then, the next values of $\boldsymbol{x}_{step}$ are calculated the matrix version of the formula (EMR). The new values of $\boldsymbol{x}_A$ are consequently calculated.

[2]Another factorization method [22]

Algorithm 1 Empirical Model Reduction function

Input: $data_{(m\times n)} = [x_1, \ldots, x_n]$
$n\_level \geq 1,\ eq\_deg \in \{1{,}2\},\ K \geq 1,\ n\_iter \geq 1,\ len_{sim} \geq 1$
Output: $sim\_data_{(len_{sim}\times n\times n\_iter)},\ R^2_{(n\_level\times n)}$

$data \leftarrow data[1:K:end]$
$max\_iter \leftarrow 50, \quad dt \leftarrow 1$
$\sigma_{data(1\times n)} \leftarrow std(data)$
$x_{(n\times \frac{m}{K}\times n\_level)}$
$x[l=1] \leftarrow data^T$
$dim_{eq\_x} = \prod_{i=1}^{n} \frac{eq\_deg+i}{i}, \quad dim_{eq\_r} \leftarrow n\_level * n + 1$
$dim_{xa} \leftarrow max(dim_{eq\_x}, dim_{eq\_r})$
$x_{A(n\_level\times dim_{xa}\times m)}, \quad A_{(n\_level\times dim_{xa})}$
$r_{(n\times n\_level)},\ \sigma_{r(n\times n\_level)}$
for $l \in \{1, \ldots, n\_level\}$ **do**
  $x[l] \leftarrow (x[l] - \mu_{x,l})/\sigma_{x,l}$
  for $t \in \{1, \ldots, m\}$ **do**
    $dx[t] \leftarrow x[t+1, l] - x[t, l]$
  end for
  if l==1 **then**
    if eq_degree==1 **then**
      $x_{Ac}[l] = [x_1, \ldots, x_n]$
    else
      $x_{Ac}[l] = [x_1, \ldots, x_n, x_1^2, \ldots, x_1x_n, \ldots x_n^2]$
    end if
  else
    $x_{Ac}[l] = [x_1, \ldots, x_n, r_{1,1}, \ldots, r_{1,n}, \ldots, r_{l-1,n}]$
  end if
  $dx' \leftarrow dx - \mu_{dx}, \quad x'_{Ac} \leftarrow x_{Ac}[l] - \mu_{x_{Ac},l}$
  $U, S, V^T \leftarrow SVD(x'_{Ac})$
  $A_c = (x'_{Ac})^{-1} \times dx' = V \times S^{-1} \times U^T \times dx'$
  $a_0 \leftarrow \mu_{dx} - A_c \times \mu_{x_{Ac},l}$
  $A[l] = [a_0, A_c],\ x_A[l] = [1, x_{Ac}]$
  $r[l] \leftarrow dx - A[l] \times x_A[l]$
  $\sigma_r[l] \leftarrow std(r[l])$
  $x[l+1] \leftarrow r[l]$
  $R^2[l] = 1 - (\sum (r[l])^2)/(\sum dx^2)$
end for

$it \leftarrow 0,\ iter\_tot \leftarrow 0$
while $it \neq n\_iter$ **do**
  $iter\_tot \leftarrow iter\_tot + 1$
  $x_{step} \leftarrow x[t = 1]$
  if $l == 1$ **then**
    if $eq\_degree == 1$ **then**
      $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}]$
    else
      $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}, x^2_{step,1}, \ldots, x_{step,1}x_{step,n}, \ldots x^2_{step,n}]$
    end if
  else
    $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}, r_{step,1,1}, \ldots, r_{step,1,n}, \ldots, r_{step,l-1,n}]$
  end if
  for $t \in \{1, \ldots, len_{sim}\}$ **do**
    $sim\_data[t, it + 1] \leftarrow x_{step}[l = 1]$
    if x_{step} diverges **then** *break*
    $r_{step}[1 : n\_level - 1] \leftarrow x_{step}[2 : n\_level] * \sigma_r[1 : n\_level - 1]$
    $r_{step}[n\_level] \leftarrow cholesky(r[n\_level]) * \sigma_r[n\_level] * rand()$
    for $l \in \{1, \ldots, n\_level\}$ **do**
      $x_{step}[l] \leftarrow x_{step}[l] + (A[l] \times x_A[l] + r_{step}[l]) * dt$
    end for
    if $l == 1$ **then**
      if $eq\_degree == 1$ **then**
        $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}]$
      else
        $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}, x^2_{step,1}, \ldots, x_{step,1}x_{step,n}, \ldots x^2_{step,n}]$
      end if
    else
      $x_A[l] = [1, x_{step,1}, \ldots, x_{step,n}, r_{step,1,1}, \ldots, r_{step,1,n}, \ldots, r_{step,l-1,n}]$
    end if
  end for
  if $sim\_data[it + 1]$ doesn't diverge **then**
    $sim\_data[it + 1] \leftarrow sim\_data[it + 1] * \sigma_{data}$
    $it \leftarrow it + 1$
  else
    if $iter\_tot \geq max\_iter$ **then** *break*
  end if
end while

Chapter 4

Used Tools and Libraries

This Chapter provides an overview of the tools and libraries used during the project. As per Chapter 3, after a brief experimentation on the MATLAB programming language, it has been decided to move the project to Python 3. Some of the main reasons were the popularity of the programming language and the presence of a well developed data ecosystem. All the experiments were conducted on Jupyter Notebooks, well known in the field of data science and machine learning.
Here follows a list of the most relevant Python libraries used in the project:

- `NumPy`: library for scientific computing, it provides a multidimensional array object, derived objects (such as matrices), and methods for fast linear algebra operations.
- `pandas`: library that provides high-performance data structures (Series and DataFrame) and operations for manipulating them.
- `SciPy`: collection of mathematical algorithms built on `NumPy`.
- `statsmodels`: library for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration [23].
- `matplotlib`: library for creating static, animated, and interactive visualizations.
- `seaborn`: data visualization library, based on `matplotlib` and integrated with `pandas` and `NumPy`, which provides a high-level interface for drawing attractive and informative statistical graphics.
- `scikit-learn`: machine learning library that provides various tools to support supervised and unsupervised learning, such as Linear Regression.

Chapter 5

The Data

This chapter examines the datasets used during the studies on the EMR model. In particular, it concerns a data exploration on both, followed by an analysis of their relative common properties. Two datasets have been used for the project:

- *Historical CMIP6 GHG Concentrations*, containing the surface mole fractions of 43 different greenhouse gases.
- *GISS Surface Temperature Analysis (GISTEMP v4)*, an estimate of global surface temperature change using temperature anomalies.

5.1 Greenhouse gases concentrations

The Historical CMIP6 GHG Concentrations dataset [24] provides measures of atmospheric concentrations (mole fractions) of the big three greenhouse gases (CO_2, CH_4, N_2O), 17 ozone-depleting substances, and 23 other fluorinated compounds. The measures were taken with dry air and reported with temporal (monthly and annual) means, as well as regional (hemisphere and global) means. These data have been made available by the Australian-German Climate & Energy College of the University of Melbourne (Australia)[1], with the purpose to be used in the Climate Model Intercomparison Project - Phase 6 (CMIP6) experiments. For the sake of simplicity, only the three main gases have been considered, analyzing their monthly mean both on a global and hemisphere scale. Also, even if data are available from year 0 (being them re-elaborated from real measures), the period of focus will be on the years 1850-2014, which better represent historical data.

[1]Dataset available at: `https://www.climatecollege.unimelb.edu.au/cmip6`

5.1.1 Data structure

Each GHG monthly mean dataset is composed of eight columns. Five of them indicating the date of the measure with multiple formats (datetime, year, month, day, and datenum). The remaining three are `data_mean_global`, `data_mean_nh`, and `data_mean_sh`, which indicate the concentration measures respectively as global, north-hemisphere, and south-hemisphere mean. The unity of measure for CO_2 is *ppm* (parts per million, corresponding to a mole fraction of $\mu mol * mol^{-1}$) while for both CH_4 and N_2O is *ppb* (parts per billion, corresponding to a mole fraction of $nmol * mol^{-1}$). In all three datasets, there are no missing values.

5.1.2 Data cleaning and processing

For a better interpretation and elaboration of data, the five temporal attributes are substituted with a singular feature `date`, having the format of `datetime64` and indicating the first day of the month of the measure. Also, the features with the format `data_mean_xxx` (with `xxx` indicating the spatial resolution of the measure) have been renominated to `ghg_mean_xxx` (with `ghg` indicating the name of the respective gas). As an example, the `data_mean_global` attribute of the CO_2 dataset has been renominated to `co2_mean_global`.

5.2 Surface temperature anomalies

The GISS Surface Temperature Analysis dataset [25, 26] provides estimates of global surface temperature change between 1881 and 2022. The dataset has been elaborated by the Goddard Institute for Space Studies of NASA, using as input two other datasets: *NOAA GHCN v4* (from meteorological stations) and *ERSST v5* (regarding ocean areas). The analysis investigates temperature anomalies rather than absolute temperatures, being the prior better subjects for the purpose (further details in the Appendix B). By temperature anomalies are indicated deviations from the normal temperature for a given location and time of year. In this case, the normal temperature is indicated as the average over the 30 years period 1951-1980 for that place and time of year.

5.2.1 Data structure

There are three initial datasets, each one referring to a different spatial resolution: global, northern hemisphere, and southern hemisphere. Each dataset is composed of 142 rows (one for each year) and 19 attributes:

- 1 attribute indicating the year

- 12 attributes (one for each month) indicating the mean temperature anomalies value for that specific month
- 4 attributes indicating the seasonal trimesters (December-January-February, March-April-May, June-July-August, September-October-November)
- 1 attribute indicating the mean value of the solar year (so considering from January to December of the same year)
- 1 attribute indicating the mean value of the so-called "meteorological year", which considers the months between December of the previous year and November of the year considered.

The unity of measure for all the measures is degree Celsius ($°C$). The only missing values are in the last row, indicating measures of the months in year 2022, year of realization of this thesis. For a better analysis of the data, the year 2022 will then not be considered.

5.2.2 Data cleaning and processing

To conform with the GHG data, only the monthly means have been considered. Also, all the three datasets have been adjusted, in order to have one row for each monthly measure. The three series have then been combined together. The final result is a single dataset similar to the GHG ones, with an attribute date for the date of the measure (the first day of the relative month) and three attributes `temp_mean_global`, `temp_mean_nh`, and `temp_mean_sh`, indicating the temperature anomalies respectively as global, north-hemisphere, and south-hemisphere mean.

5.3 Data analysis techniques

This section describes some of the techniques used in the analysis of the data series, both in the case of the GHG dataset and the temperature anomalies dataset.

5.3.1 Fourier transform of time series

Given a time series $x_1, ..., x_n$ its **Discrete Fourier Transform (DFT)** can be defined as:

$$d(\omega_k) = n^{-1/2} \sum_{t=1}^{n} x_t e^{2\pi i \omega_k t}$$

for $k = 0, 1, ..., n-1$, where the frequencies are called the Fourier or fundamental frequencies [27]. In this way, the time-dependent function describing the time series is transformed into a frequency-dependent function. Each of the values in the

outcome series indicates the strength of the corresponding fundamental frequency and they can also be remapped as the strength of specific periods (simply remapping from the frequencies using $T = 1/f$). Understanding the fundamental frequencies can help to better grasp the seasonalities present in the series (further details in the next section). In this particular case, it has been used the `fft` function of `numpy`, which applies a particular type of DFT called Fast Fourier Transform (FFT) [28].

5.3.2 STL decomposition

Any time series function can be divided into three core components:

- The trend T, which shows the movement of the series over a long period of time.
- The seasonality S, which describes the presence of variations that occur at specific regular intervals and shows a repeating short-term cycle in the series.
- The noise N, which is the remaining random variation in the series.

Considering our original time series X as an additive model, it is possible to describe the decomposition as

$$X[t] = T[t] + S[t] + N[t]$$

with $t = 1, ..., n$. One of the main decomposition approaches used is called **STL (Seasonal Trend decomposition based on Loess)** [29], which can be exploited in Python 3 thanks to the `stl` function of `statsmodels`.

5.3.3 Autocorrelation and partial autocorrelation of time series

Correlation is a statistical relationship between two random variables or, in this particular case, between two time series. One of the most used measures for linear correlation is the Pearson correlation coefficient which, given two random variables X and Y, can be defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where $cov(X,Y)$ is the covariance of X and Y and σ_X and σ_Y are the standard deviations of the two random variables. The Pearson correlation coefficient values range between -1 and 1, with:

- $\rho_{X,Y} = 1$ indicating a perfect positive correlation between X and Y (when one increases, so does the other).

- $\rho_{X,Y} = -1$ indicating a perfect negative correlation between X and Y (when one increases, the other decreases).
- $\rho_{X,Y} = 0$ indicating no linear dependency between X and Y.

In the case of time series it is not only possible to study the correlation between different time series, but also between different lagged versions of a singular one with its original. In this particular case, we talk about Autocorrelation. It is often studied analyzing the Autocorrelation Function (ACF), which calculates the autocorrelations in function of the lag k, taking into account also the previous lags between 0 and $k-1$. On the other hand, with the Partial Autocorrelation Function (PACF), only the correlation with the k-lagged version is taken into account. In this way, it is possible to better estimate the correlations between lagged versions of the series and verify the presence of seasonalities in the data.

5.4 Data analysis results

This section reports the results of the data analysis. Some of the figures cited can be found in Appendix C. The figures also represent the data as described in the legend.

5.4.1 Carbon dioxide

As a first step, it is possible to plot the three time series to have a general idea of the behavior of the data. It can be observed that the series overlap, with the trends that tend to rise. On the other hand, when zooming on a small range of dates it is possible to observe a variation that seems periodical and that suggests periodicity in the time series.

In order to deepen this theory, the Fourier transforms of the series are examined. Figure 5.2 shows that the transforms of the global means and the northern hemisphere means have similar behaviors, with comparable peaks on the same fundamental frequencies (or periods). In particular, the most evident is the one at period$\approx$12, showing an annual periodicity which will be further examined later. On the other hand, the southern hemisphere data show a particular behavior, having the 12-months peak replaced by a smaller 6-months one.

Once established the presence of periodicities, it is possible to extract some more information from the time series, decomposing them. For the sake of simplicity, all three time series will be studied on cycles of 12 months. The seasonality (Figure 5.3) of the global and north-hemisphere data shows a minimum value in August and a maximum in April. On the other hand, the southern region appears to be more irregular, confirming the absence of a steady annual cycle. It must be noted that the amplitude of the curve in the north-hemisphere data is far bigger than the

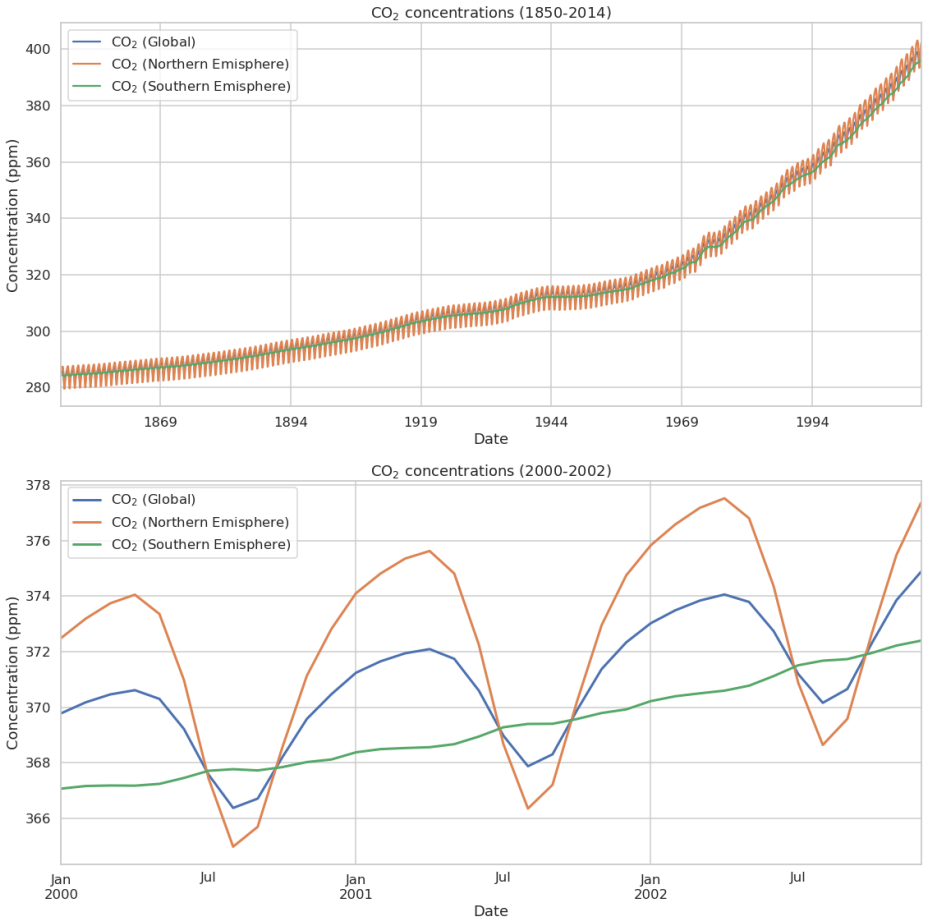


Figure 5.1: CO_2 concentrations evolution (top) between 1850 and 2014 and (bottom) between 2000 and 2002.

one in the south-hemisphere data. This demonstrates a much greater contribution of the northern region to the global CO_2 cyclicity, most likely due to a greater presence of industrially developed nations. Examining the trends (Figure C.1), all the three time series show an increase over time, confirming the initial evaluations of the data behavior. As evident also in the analysis of the other two gases, the noise component (Figure C.2) presents an unusual behavior: before a certain year (1959 in the case of CO_2) the amplitude tends to be really lower than afterwards. This is probably caused by the fact that the dataset contains older data that passed through an elaboration, which might have been more intense.

One last thing which might be interesting to examine is the partial autocorrelation of the data. From Figure C.3 it is possible to observe again the 12-months periodicity in the behavior of the correlation, with the amplitudes which slowly

decrease due to the growing trend.

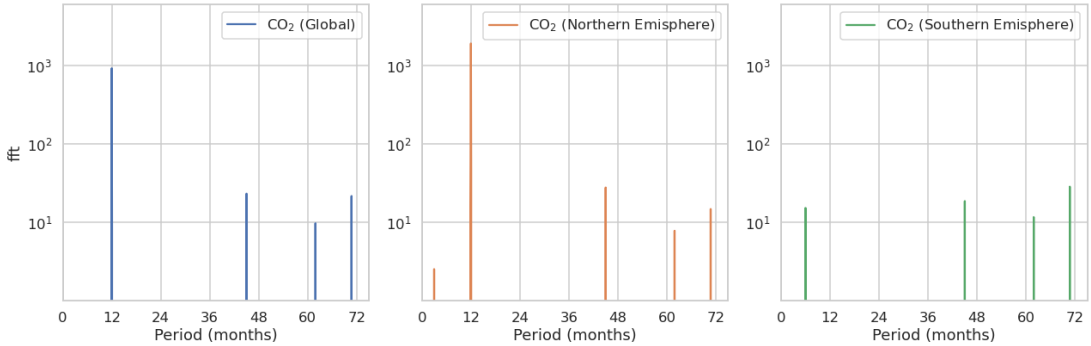


Figure 5.2: Fourier transforms of (from left to right) global, northern hemisphere and southern hemisphere concentrations of CO_2 in function of period.

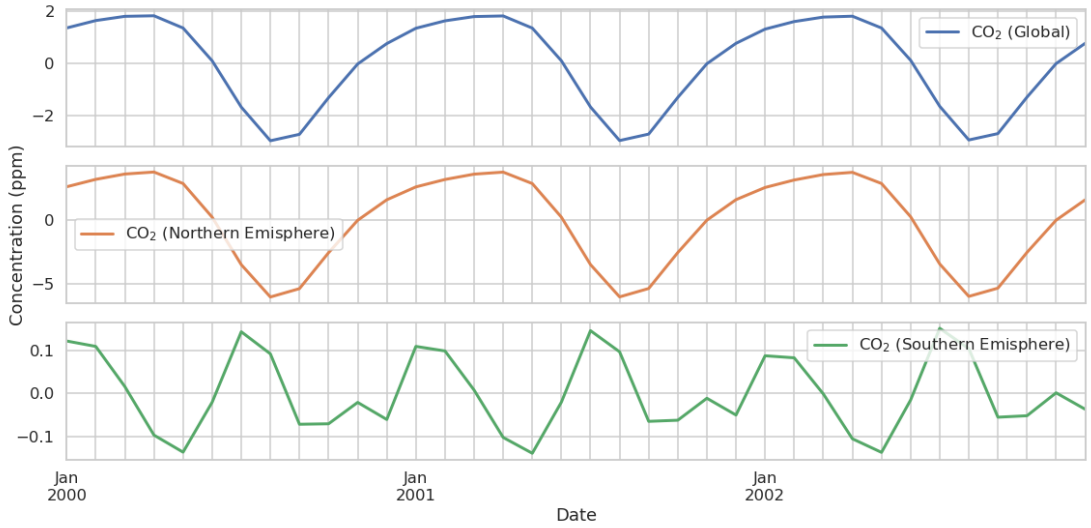


Figure 5.3: Seasonality components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CO_2 between 2000 and 2002.

5.4.2 Methane

As for the CO_2 analysis, the best point to start is by directly plotting the time series (Figure 5.4). It is immediately noticeable how the three series are easily distinguishable. The northern hemisphere data are far higher than the southern hemisphere ones. This can be attributed once again to a greater presence of developed nations in the north, which brings not only greater use of methane-based heating systems but also more emissions due to intensive agriculture and farming.

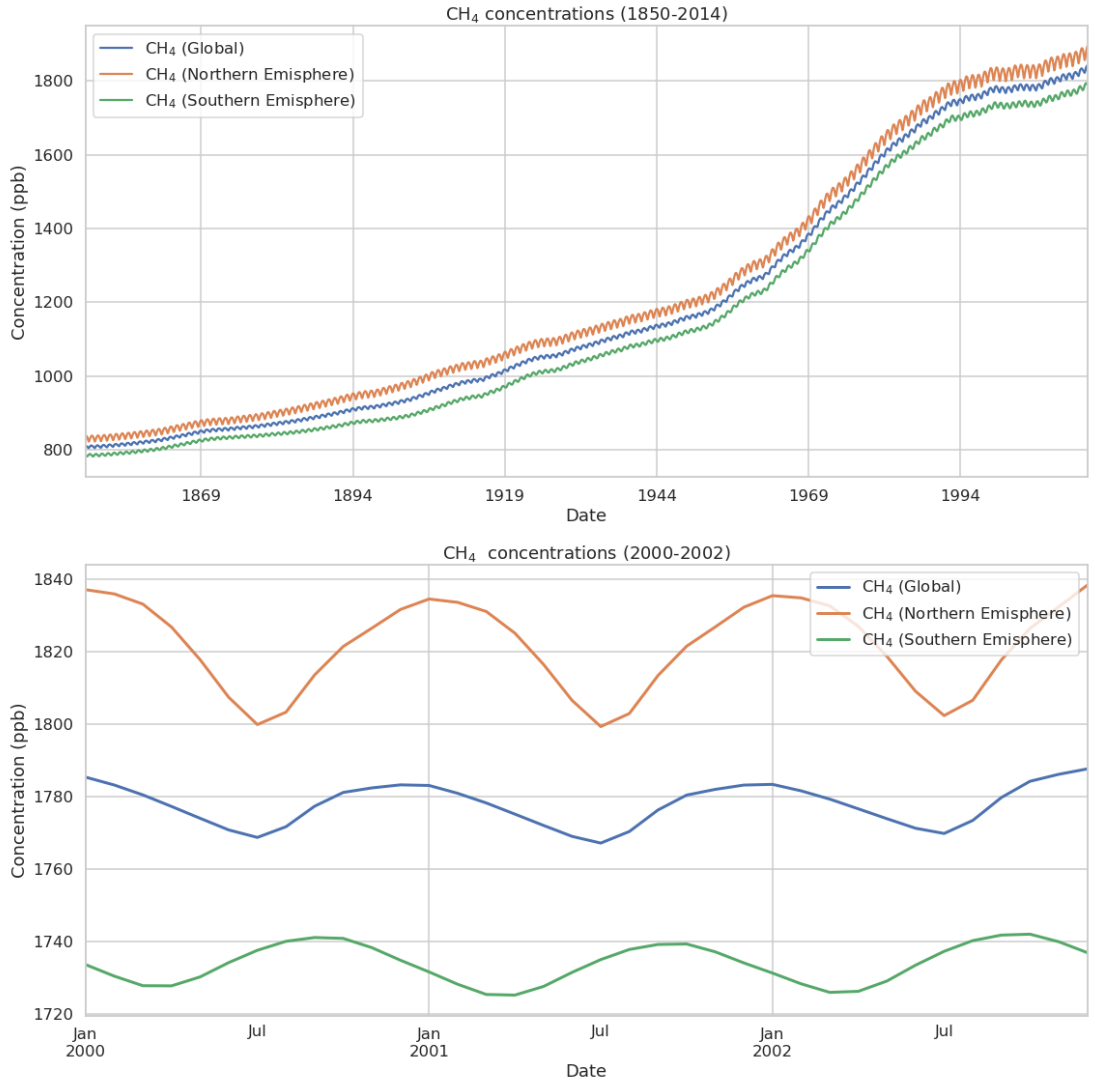


Figure 5.4: CH_4 concentrations evolution (top) between 1850 and 2014 and (bottom) between 2000 and 2002.

By analyzing the Fourier transforms (Figure 5.5), it is possible again to confirm the presence of a 12-months periodicity, this time also with the southern hemisphere data. There are also other periodicities noticeable from the transformations of the three series, above all one with period≈180 months (15 years), with high peaks in all the three graphs.

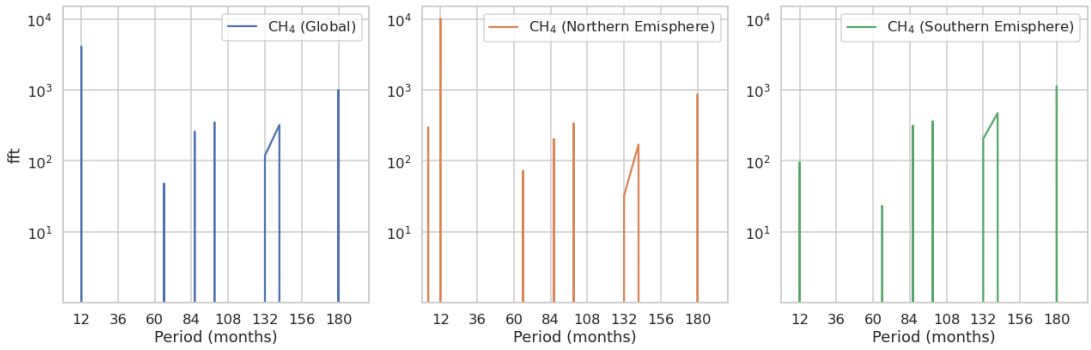


Figure 5.5: Fourier transforms of (from left to right) global, northern hemisphere and southern hemisphere concentrations of CH_4 in function of period.

Using the 12 months period, it is possible to extract again the seasonality, the trend, and the noise of the time series with the STL approach. The seasonality (Figure 5.6) in the global and northern hemisphere data shows similar performances, with a minimum in July and a maximum in January. This behavior, at least in the northern hemisphere, might be compatible with the use of methane-based heating systems during colder periods. The southern hemisphere data show a similar but delayed behavior, with the peaks at September-October and the minimum at March-April. The trends, as usual, are rising (Figure C.4).
Finally, the partial autocorrelation (Figure C.4) shows again a 12-months periodicity.

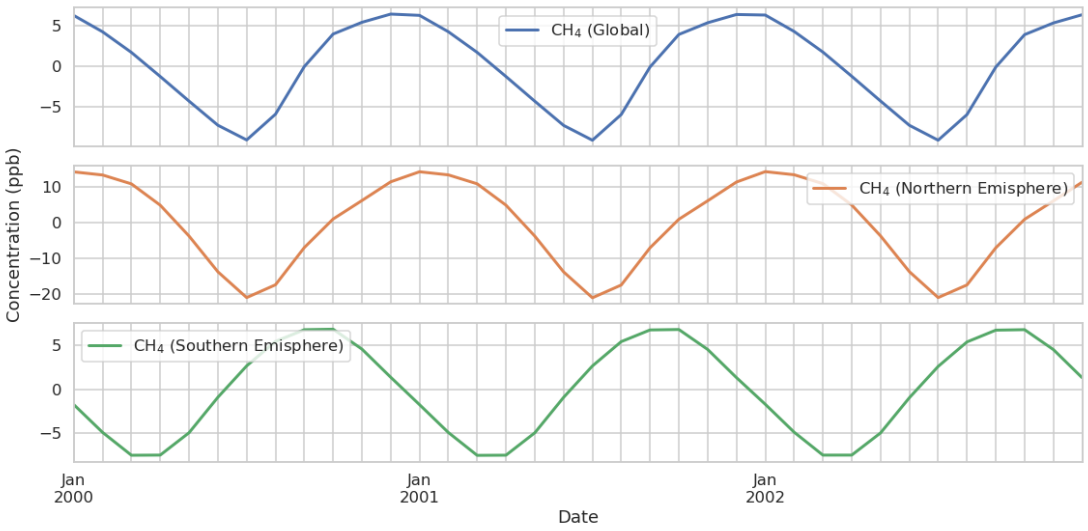


Figure 5.6: Seasonality components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CH_4 between 2000 and 2002.

5.4.3 Nitrous oxide

Figure 5.7 shows how, similarly to the CH_4 case, the three time series are distinguishable, even if not as much as in the previous case. The northern hemisphere data values are almost always higher than the southern hemisphere ones by 1 ppb.

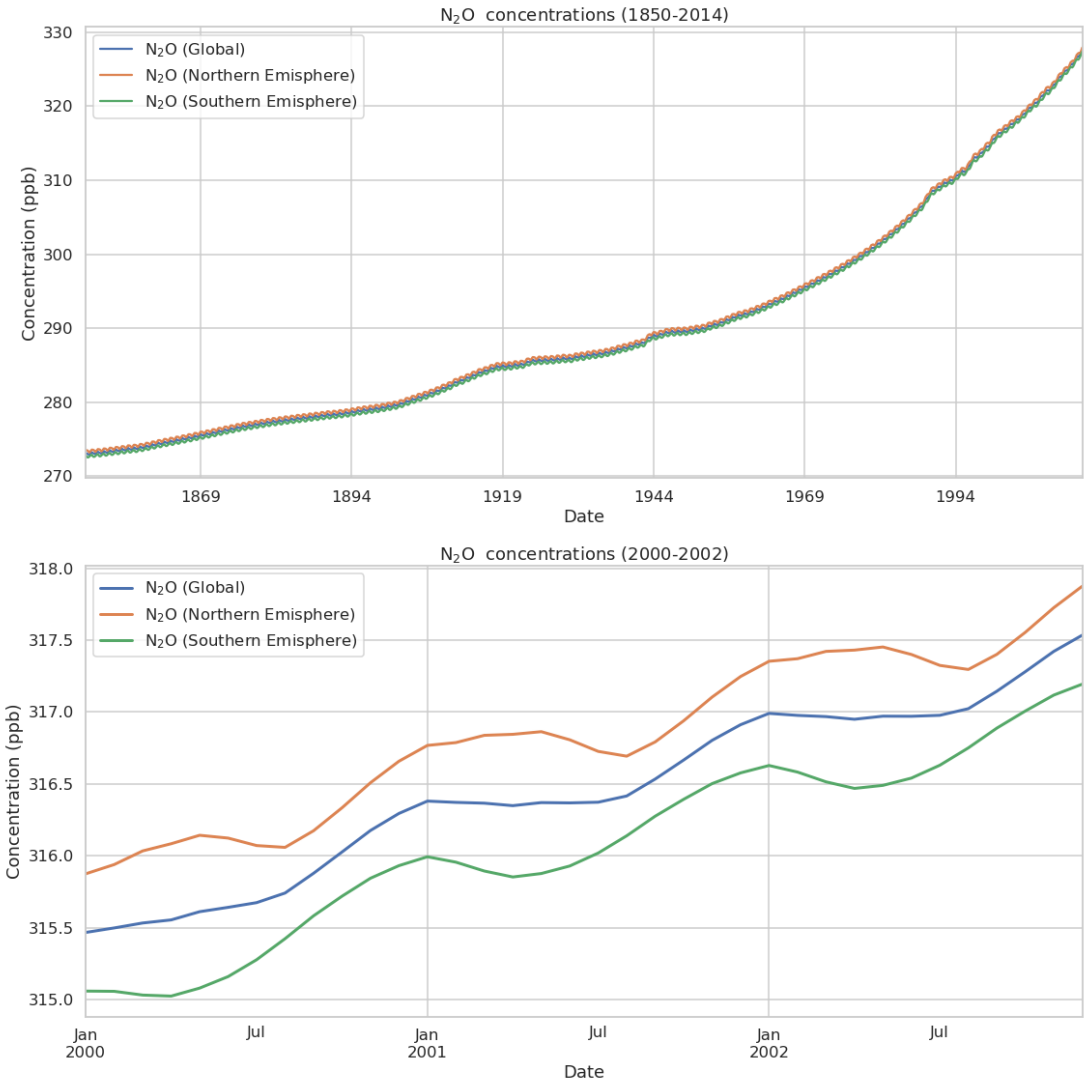


Figure 5.7: N_2O concentrations evolution (top) between 1850 and 2014 and (bottom) between 2000 and 2002.

The Fourier transforms (Figure 5.8) don't show many frequency peaks, but only the usual one with period≈12 months, with all the three time series.
Applying again the STL approach over a 12 months period, the series is decomposed

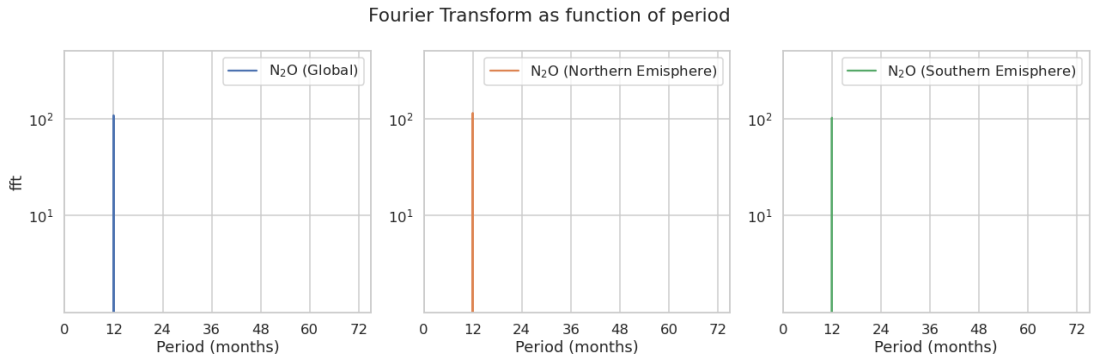


Figure 5.8: Fourier transforms of (from left to right) global, northern hemisphere and southern hemisphere concentrations of N_2O in function of period.

into seasonality, trend, and noise. This time, all three series have the maximum seasonality value (Figure 5.9) in January. On the other hand, global and northern hemisphere seasonalities have a minimum in August, while the southern hemisphere one in June. The trends (Figure C.7) are all rising with the same curve, with the northern hemisphere values slightly higher than the southern hemisphere ones, as already shown in the first plot.

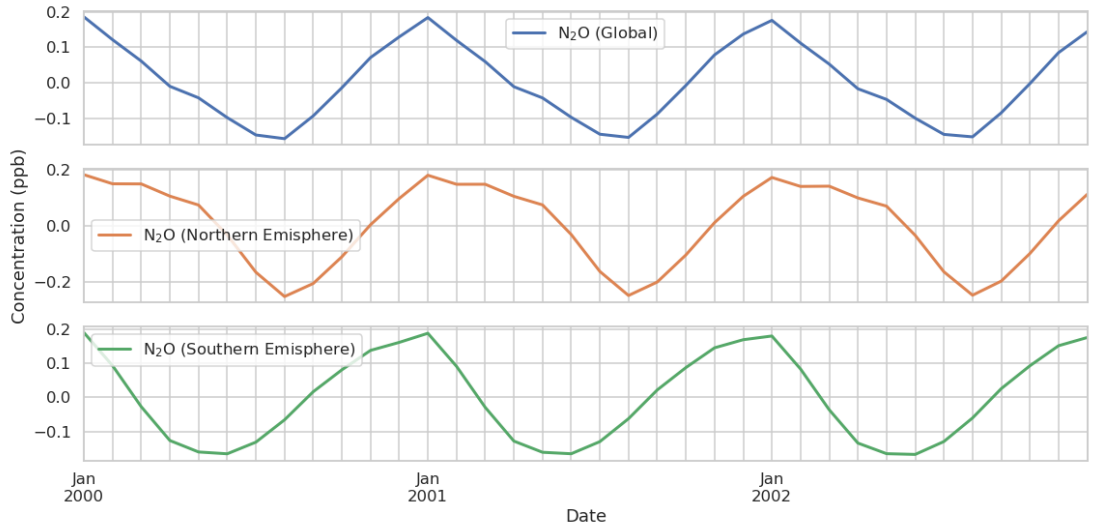


Figure 5.9: Seasonality components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of N_2O between 2000 and 2002.

The partial autocorrelation (Figure C.9) confirms again the 12-months periodicity.

5.4.4 Temperature anomalies

In Figure 5.10 it is possible to observe how the initial data is considerably more variable than the GHG ones. Anyway, it is obvious that the trend is rising in all three cases.

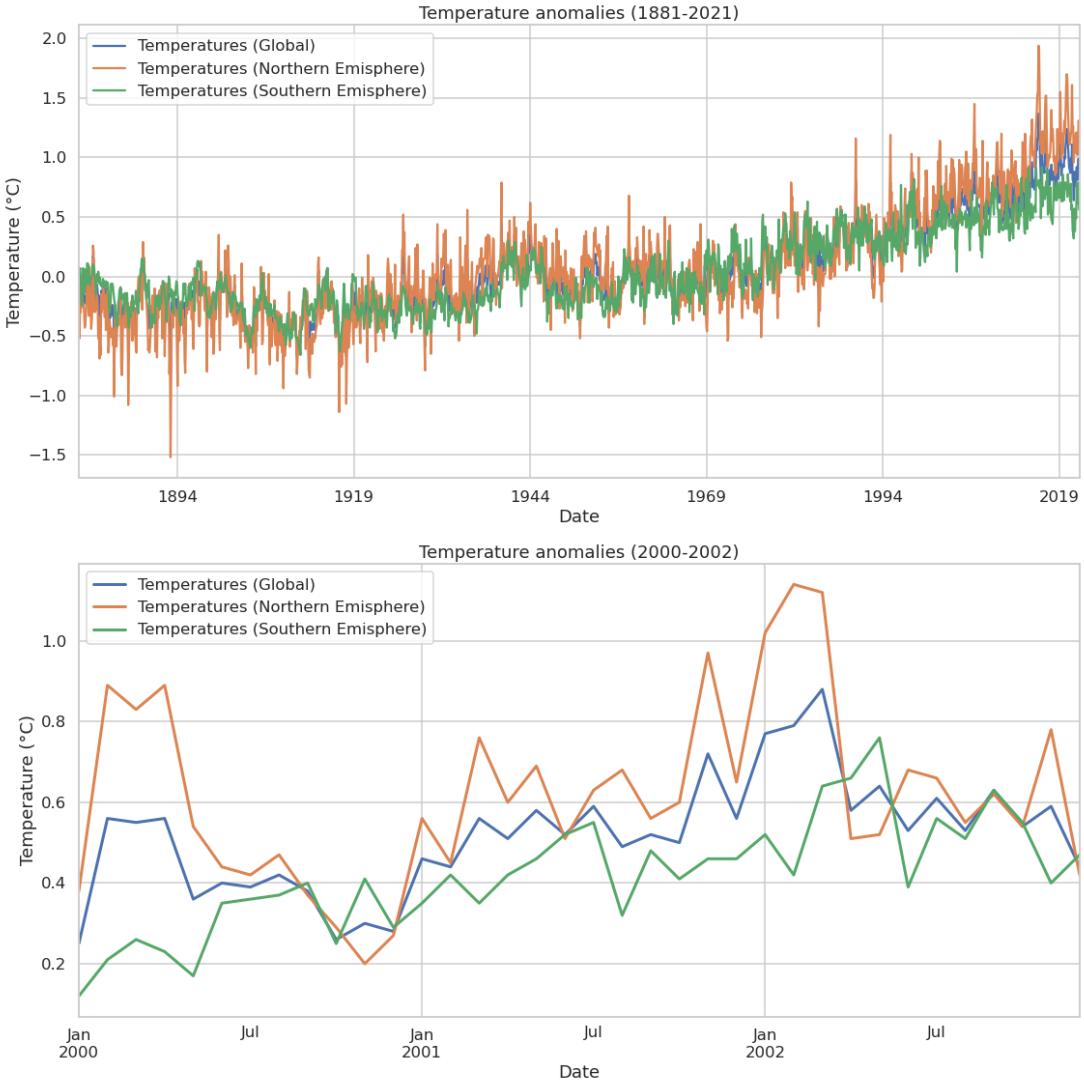


Figure 5.10: Temperature anomalies evolution (top) between 1881 and 2021 and (bottom) between 2000 and 2002.

Also, the Fourier transforms of the data give significantly different results (Figure 5.11). In fact, numerous peaks are present, spread over various periods. In order to uniform and better confront the data, also in this case the analysis will be

conducted over a 12 months period, even if it doesn't seem to be a particularly relevant period compared to the others.

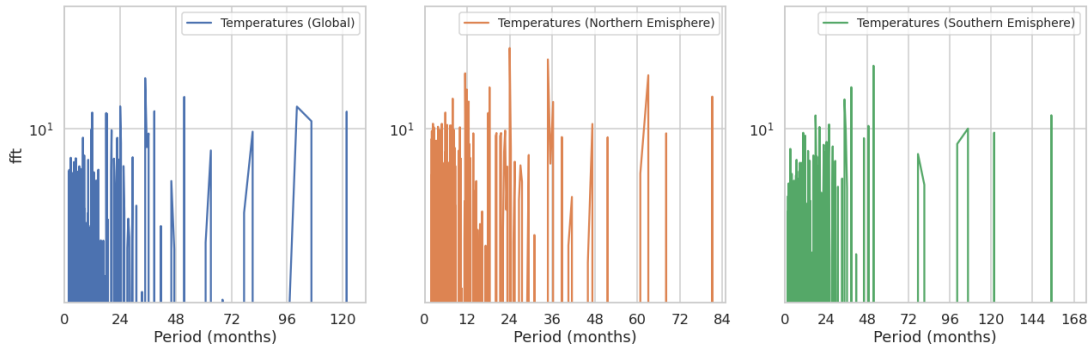


Figure 5.11: Fourier transforms of (from left to right) global, northern hemisphere and southern hemisphere temperature anomalies in function of period.

Using the STL decomposition, the seasonality, trend, and noise over a 12 months periodicity are obtained. The seasonality (Figure 5.12), from a general overview, doesn't seem to be constant. Anyway, a particular cycle behavior is present at the global and northern hemisphere level: all the maximum peaks seem to be in the months between October and March while having more stable data in the other months. The trends of the three series (Figure C.10), even if less stable, tend to rise, confirming the initial observation of the data. Finally, the noise seems to correspond to a white one (Figure C.11).

The partial autocorrelation (Figure C.12) seems this time to be more variable with respect to the lags, even with what seems to be a pattern in the alternation of positive and negative correlations.

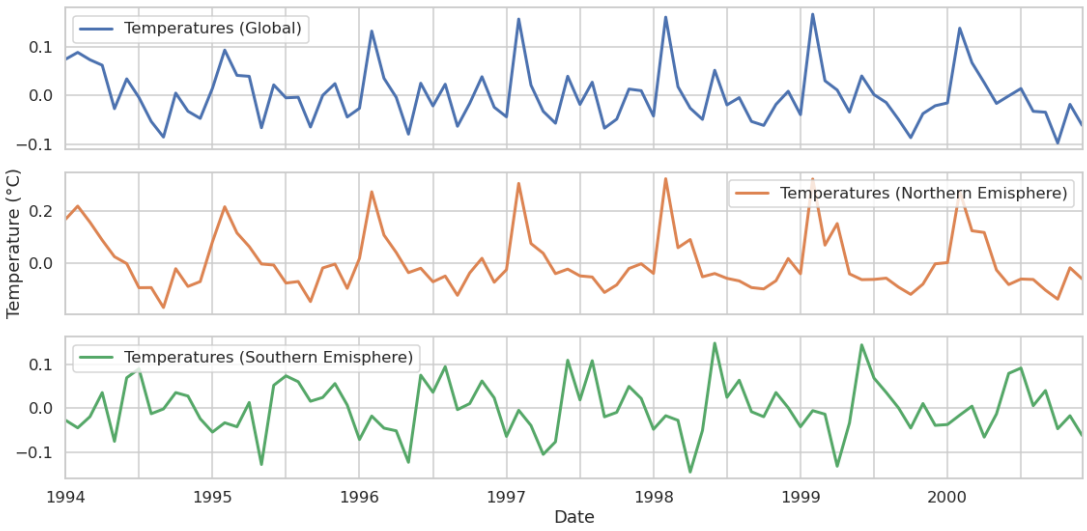


Figure 5.12: Seasonality components of (from top to bottom) global, northern hemisphere and southern hemisphere temperature anomalies between 1994 and 2002.

5.4.5 Cross-analysis

After having studied the three GHG and the temperature anomalies data separately, it is important to analyze the cross-correlations between them. In this particular case, they are divided into two groups, global means and northern-southern hemisphere means, confronting them using the Pearson correlation coefficient. In order to achieve these results, the GHG and temperature anomalies datasets have been merged, with an inner join over the dates. The new dataset have 13 attributes (one for dates and 3 spatial means for each measure) and 1608 rows, with a temporal range between 1881 and 2014. From a first observation of the global means data (in Figure 5.13), the high correlation between the three GHG is immediately visible. All three of them in fact have a coefficient greater than 0.95. On the other hand, all three of them seem to have in any case a strong positive correlation with the temperature anomalies, probably due to the rising trend in all four series. To gain a deeper knowledge of the mutual relation between the data, it's possible to confront the correlations of the data seasonalities, obtaining them once again with the STL decomposition. The only relevant information seems to be a strong positive correlation between N_2O and CH_4 gases, suggesting similar yearly behaviors.

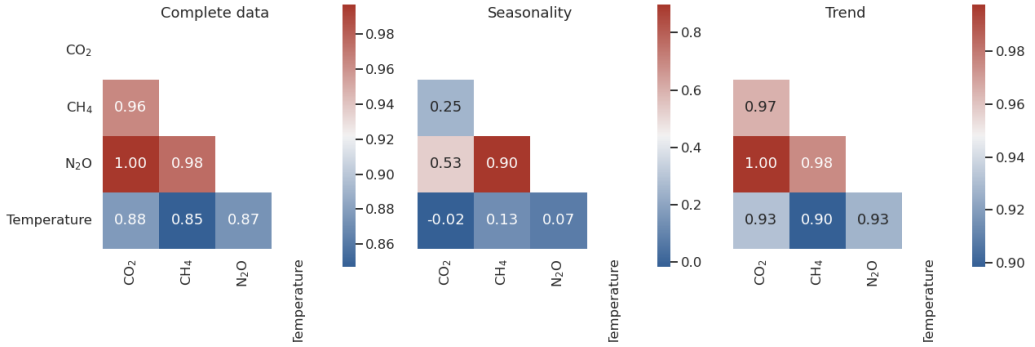


Figure 5.13: Correlation heatmaps of global data.

After having examined the global means, more detailed relations can be studied using the northern and southern hemisphere means of all the data. Both the correlations between the three gases and the gases with temperature anomalies show to be strongly positive. A particular data that stands out is the less strong correlation between CH_4 levels and temperature anomalies in the northern hemisphere, suggesting a lower contribution of the gas in the temperature rising when compared to the others. Finally, the seasonalities study shows strong positive correlations between the three gases behavior in the northern hemisphere, indicating similar yearly behaviors. Also, some particular results can be found in the strong negative correlations between the CH_4 seasonality in the southern hemisphere and the other two gases seasonalities in the northern one.

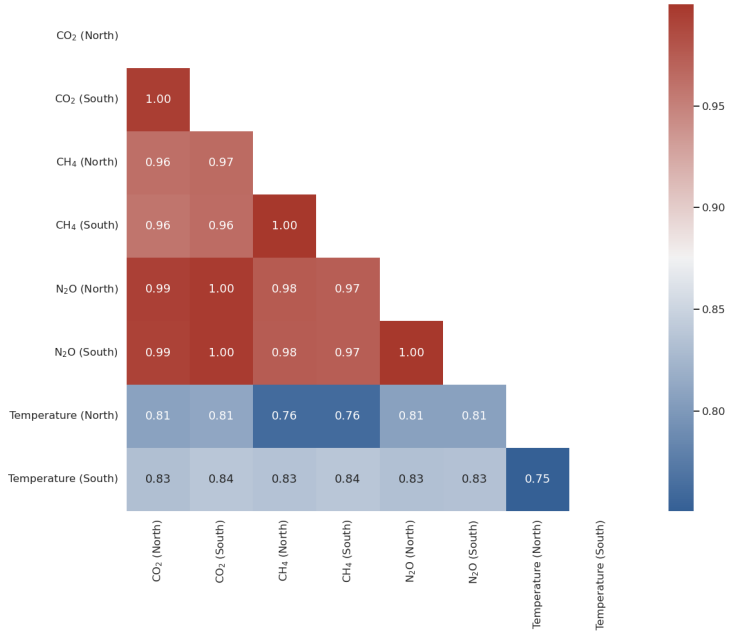


Figure 5.14: Correlation heatmaps of hemispheres data.

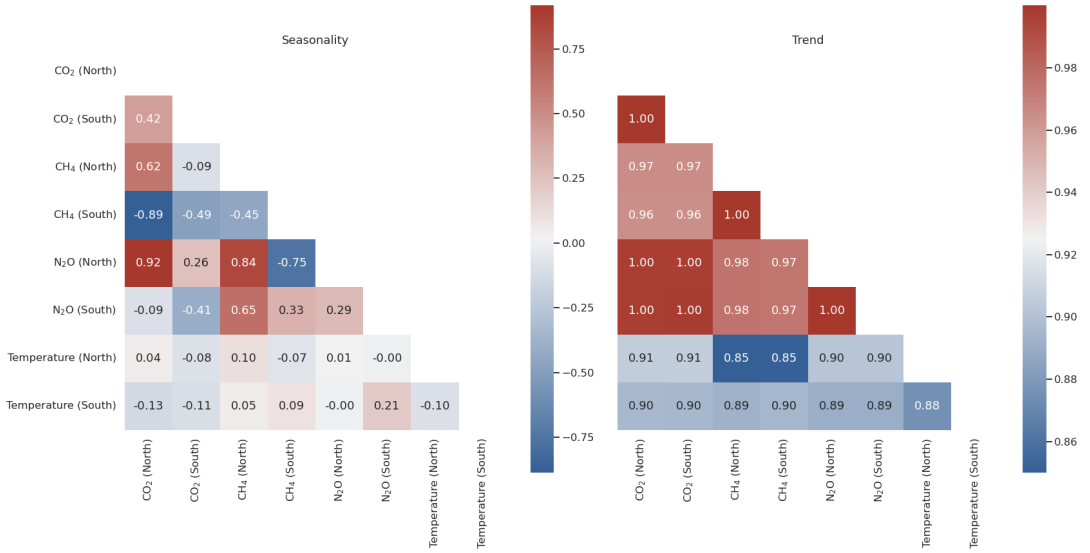


Figure 5.15: Correlation heatmaps of hemispheres data (seasonalities and trends).

Chapter 6

Empirical Model Reduction: Experiments and Results

This chapter provides the results obtained by applying the empirical model reduction method to various versions of the dataset described in Chapter 5. First of all, the tuning and analysis pipeline, common to all the cases analyzed, will be presented. The results from the singular data cases analyzed will be then discussed, verifying the best models for each one of them. For the sake of simplicity, only the global data of the greenhouse gases and temperature anomalies have been considered.

It is important to note that, in all the cases, the data were derived by a detrended version of the original one, using the STL decomposition (discussed in Chapter 5) to calculate the trends and subtracting them from the original time series. In this way, the system presents itself as a stochastic process, ideal for an EMR study.

6.1 The model tuning

As presented in Chapter 3, the EMR function comes with a series of hyperparameters (like the number of levels of the model) which are fundamental requirements in order to determine how the algorithm should work. Clearly, varying the parameters influences the performance and the resulting simulated data.

In machine learning, hyperparameter tuning is a well-known type of optimization problem, where different combinations of hyperparameters are evaluated in order to find the set which gives the most performing model. Since the EMR method works in a similar way, it is possible to study the performance of the models realized by varying the values of the hyperparameters. In this case, a method called grid search has been used, by which a series of possible values for the hyperparameters are used to train models, evaluating the best ones. The hyperparameters subject

of study were:

- The coefficient of sub-sampling K, using as possible values 1, 3 and 6 (which correspond to considering all the months, one for trimester and one for semester).
- The degree of the main level equation eq_deg, considering linear ($eq\_deg = 1$) and quadratic ($eq\_deg = 2$) EMR models (experiments have shown that, with $eq\_deg > 2$, it's not possible to construct a stable model based on the climate data).
- The number of levels n_level, with values between 1 and 10, studying the dynamics of a model with up to 9 residual levels.

The length of the reconstructed data has been kept equal to the length of the original ones (considering the sub-sampling). Also, for all the possible combinations of hyperparameters, 10 simulations were performed.

The R^2 coefficient has been chosen as the evaluation criteria. In particular, since the model better performs if the coefficient of determination of all the components is close to 0.5, the cases with the mean of R^2 over the components close to 0.5 and the lowest standard deviation are considered as the best performing ones. However, as shown in Appendix A, a coefficient $R^2 \approx 0.5$ doesn't always bring to a perfect representation of the data. It only indicates the optimal number of levels needed to build a well performing model. So, in order to better evaluate the results, some promising cases will be selected for a deeper examination, analyzing the simulated data themselves, their autocorrelation function, and the probability density function compared to the original ones.

6.2 Original time series

The first analysis to begin with regards the EMR models constructed using the original time series. Except for the de-trending, no further modifications were made.

6.2.1 Results

Table 6.1 reports the results of the model tuning using the original time series. In particular, the mean and standard deviation values of the R^2 coefficient over each component are reported. Where indicated "x" it means that the data produced by the model diverged during at least one of the ten simulations.

Generally, the table presents results with R^2 mean near 0.5 and a really low standard deviation. Some particular cases will be better examined in this chapter, studying the autocorrelation and probability density function of the simulations.

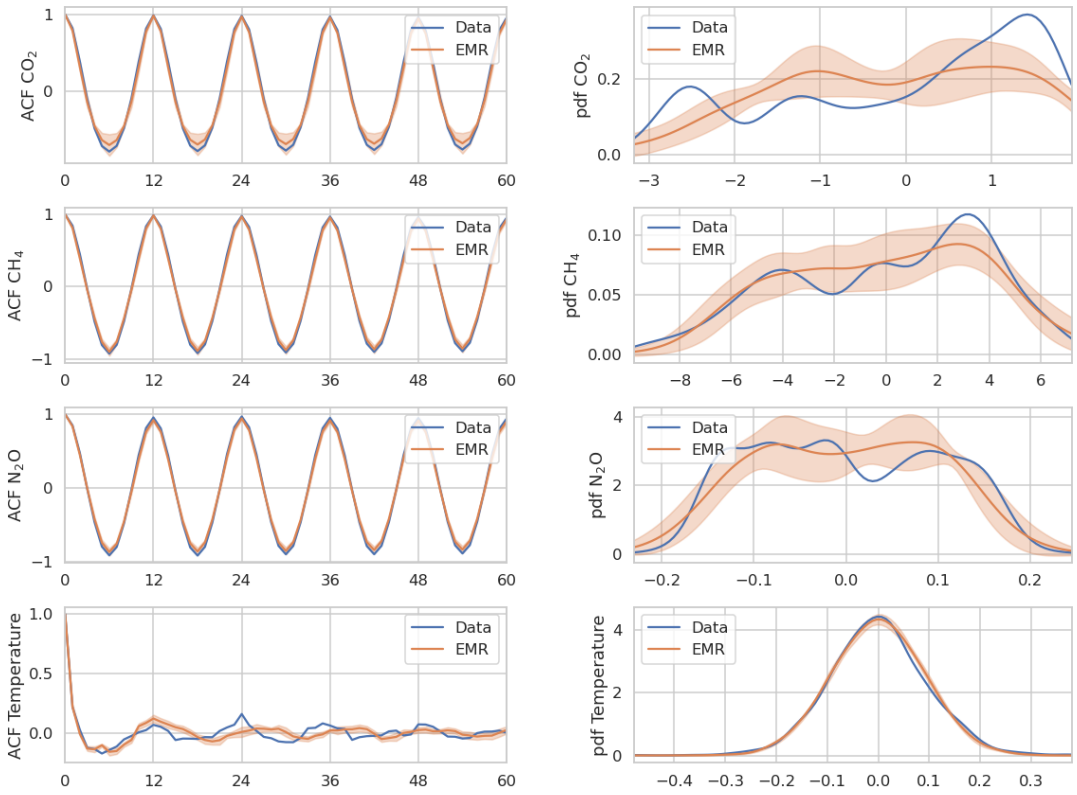
It is important to observe how, increasing the value of K, decreases the maximum number of levels accepted before a divergence. It is finally interesting to note that the only way to build a quadratic EMR model with this dataset is through a heavy sub-sampling, with $K = 6$. Even in this case, the system was able to build stable models only with a low quantity of levels.

	$K = 1$				$K = 3$				$K = 6$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.74	0.20	x	x	0.88	0.18	x	x	0.90	0.16	0.91	0.15
2	0.52	0.10	x	x	0.63	0.14	x	x	0.72	0.17	0.54	0.04
3	0.73	0.13	x	x	0.65	0.10	x	x	0.56	0.10	x	x
4	0.56	0.04	x	x	0.57	0.07	x	x	**0.53**	**0.08**	0.54	0.02
5	0.58	0.04	x	x	0.63	0.08	x	x	x	x	**0.52**	**0.01**
6	0.70	0.12	x	x	*0.50*	*0.08*	x	x	x	x	x	x
7	0.53	0.03	x	x	0.49	0.04	x	x	x	x	x	x
8	0.52	0.02	x	x	**0.52**	**0.02**	x	x	x	x	x	x
9	**0.52**	**0.01**	x	x	0.54	0.04	x	x	x	x	x	x
10	0.61	0.06	x	x	x	x	x	x	x	x	x	x

Table 6.1: R^2 results applying the original dataset. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $K = 1$, linear EMR, $n\_level = 9$

This was the best case $K = 1$ and a linear EMR, according to the established metrics.
It is already visible in Figure 6.1(a) how the autocorrelation function behaves better than in the previous case, keeping a steady correlation with the greenhouse gases and behavior with the temperature more similar to the original one. The pdfs shown in Figure 6.1(b) better cover the value distribution of the original data. Figure 6.2 shows how the simulated GHGs, perfectly fitting initially, manage to keep a similar behavior over time, although increasing the standard deviation of the data. Really peculiar is the behavior of the temperature: it perfectly overlaps the original data for nine months, but afterwards it is not able to well represent the data dynamics anymore.



(a) ACF (b) pdf

Figure 6.1: Autocorrelation function (a) and probability density function (b) of the case with $K = 1$, linear EMR, and $n\_level = 9$.

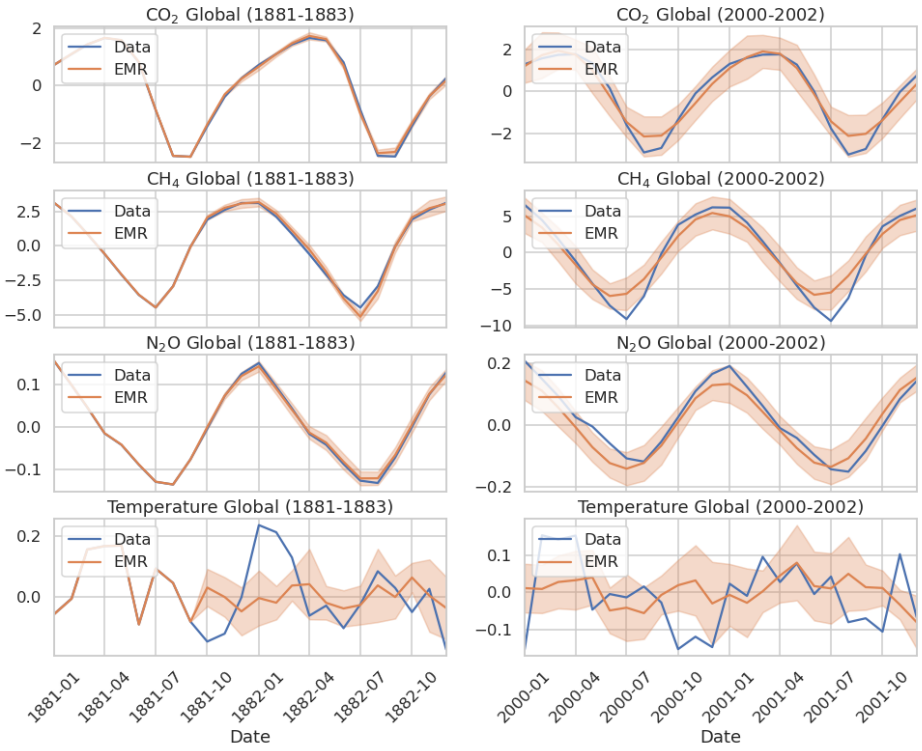
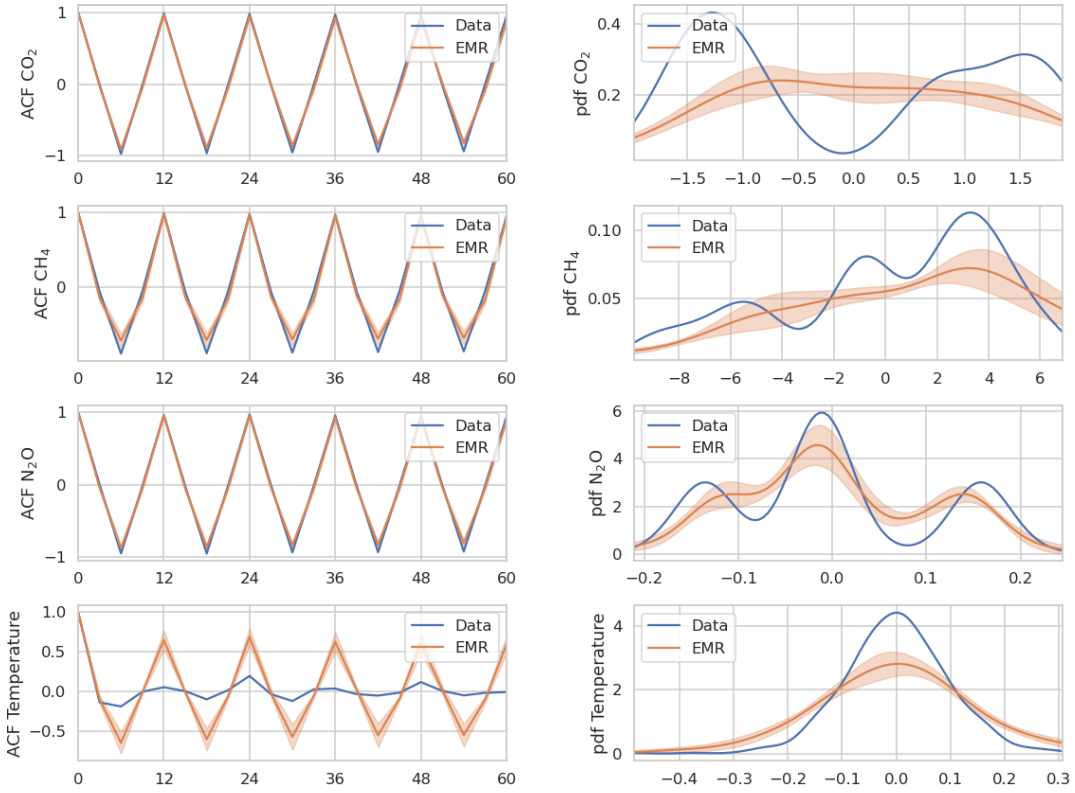


Figure 6.2: Comparison between the original data and the simulation of the case with $K = 1$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $K = 3$, linear EMR, $n\_level = 8$

This case presents an acceptable mean R^2 value and a really low standard deviation. For these reasons it has been taken into account.
Except for the case of temperature anomalies, the autocorrelation functions (Figure 6.3(a)) seem to be perfectly represented. On the other hand, the probability density functions (Figure 6.3(b)) show poor results, with all the components not well represented.
The consequences can be seen in Figure 6.4 where the simulation doesn't perfectly describe the data, even if performing well initially.
In general, it is possible to agree that this case is not good as the previous one, with $n\_level = 6$.



(a) ACF (b) pdf

Figure 6.3: Autocorrelation function (a) and probability density function (b) of the case with $K = 3$, linear EMR, and $n\_level = 8$.

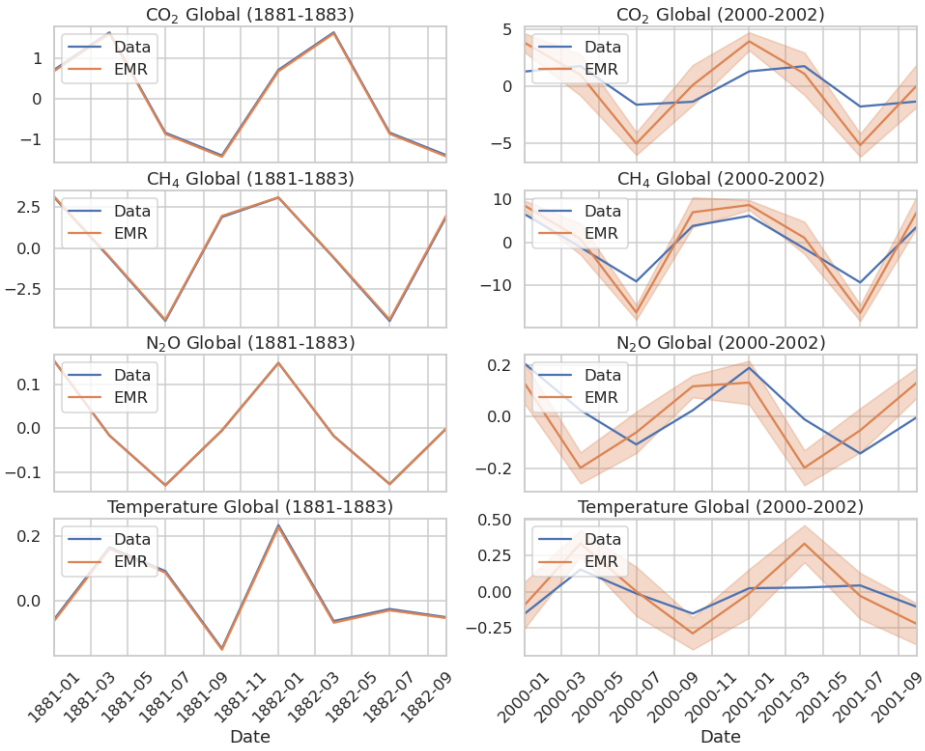
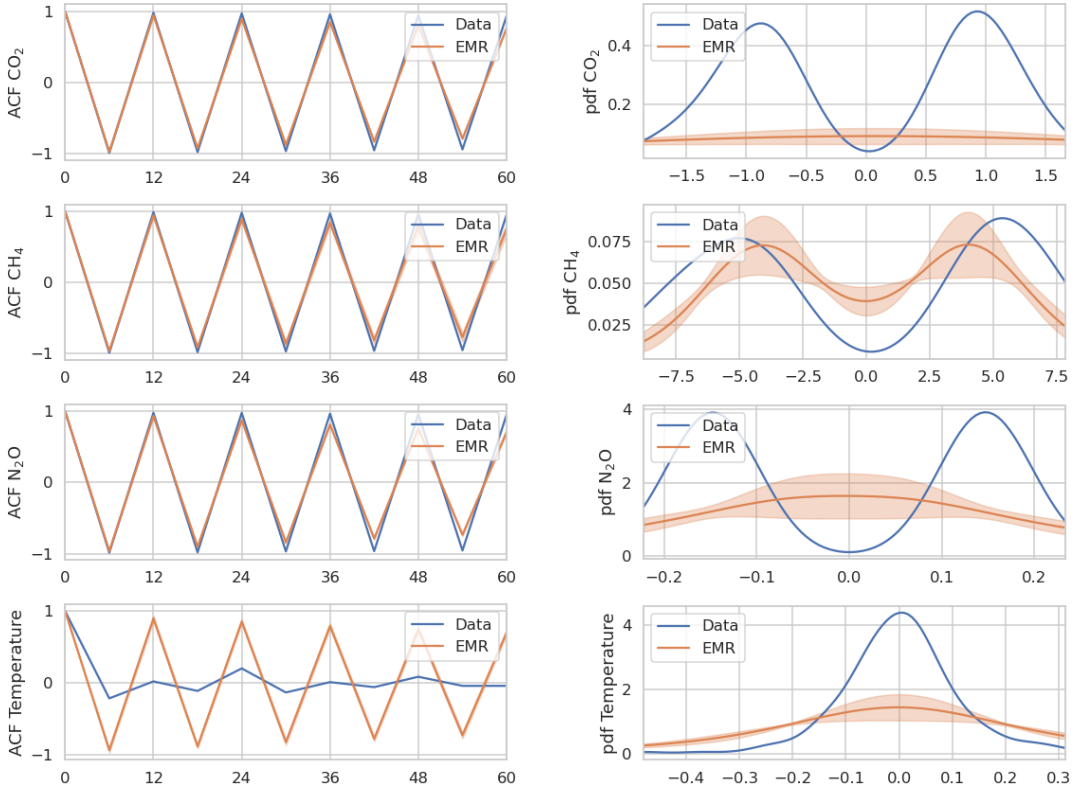


Figure 6.4: Comparison between the original data and the simulation of the case with $K = 3$, linear EMR, and $n\_level = 8$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $K = 6$, linear EMR, $n\_level = 4$

With a sub-sampling of one measure every semester, it is surely more difficult to well represent the data. This can be already observed in Figure 6.5(a), with the autocorrelations that slowly tend to reduce themselves increasing the lag. Also, it is important to notice that the autocorrelation function of the temperature anomalies is poorly represented. The probability density functions (Figure 6.5(b)) show even worse results.
These factors are clear signs of a bad reproduction of the data. Figure 6.6 shows how the simulated signals tend to de-correlate over time.



(a) ACF (b) pdf

Figure 6.5: Autocorrelation function (a) and probability density function (b) of the case with $K = 6$, linear EMR, and $n\_level = 4$.

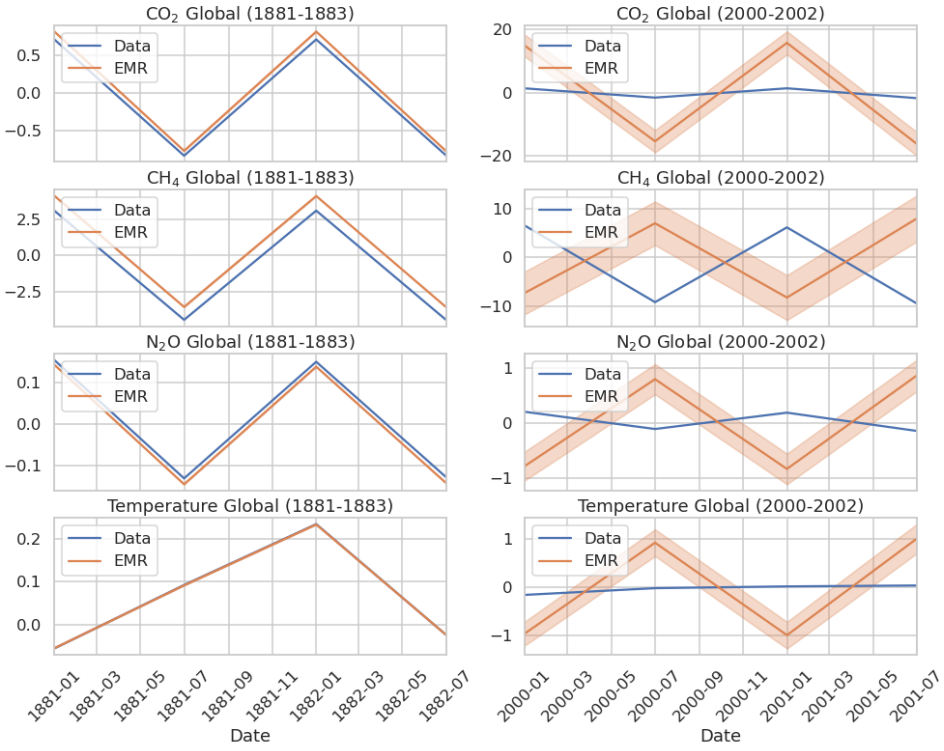
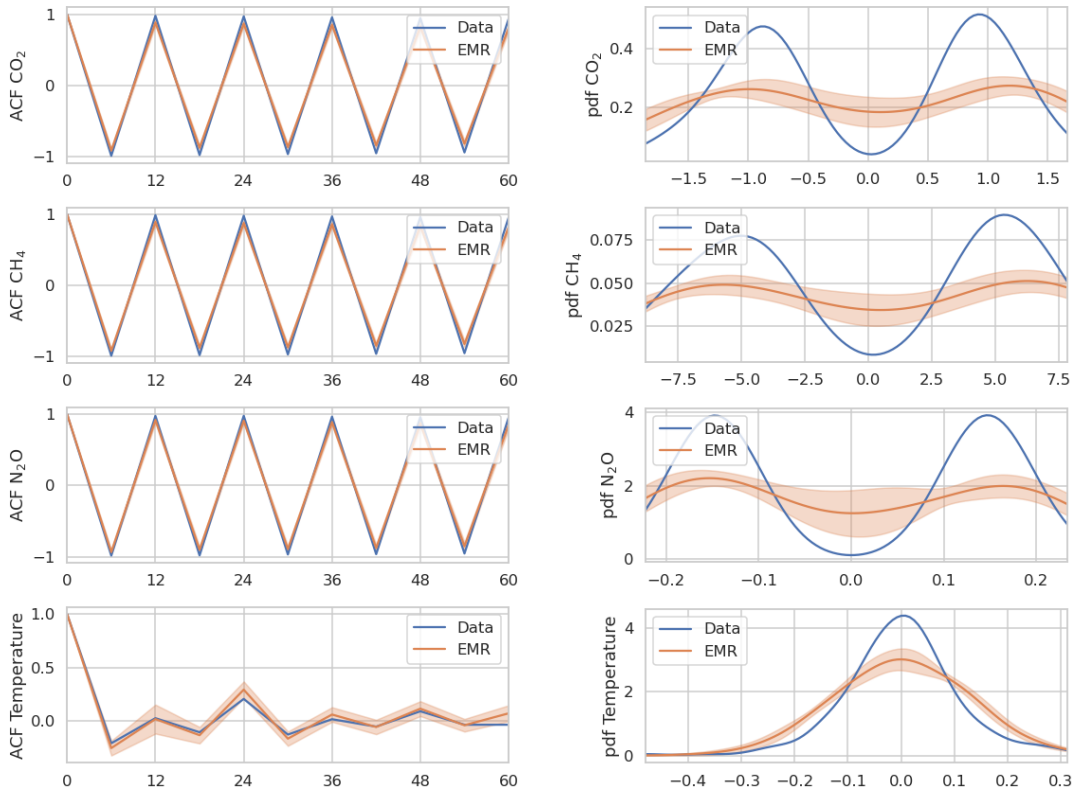


Figure 6.6: Comparison between the original data and the simulation of the case with $K = 6$, linear EMR, and $n\_level = 4$.

Case $K = 6$, quadratic EMR, $n\_level = 5$

Being one of the few cases with a quadratic EMR, this surely is an interesting subject of study.
While still having a bad simulation of the values distributions, Figure 6.7(a) shows that in this case the autocorrelation functions were perfectly represented, even with the temperature measures.
This brings to a good simulation of the data (Figure 6.8), although with some difficulties in maintaining it similar to the original over a long period of time.
In general, this case showed to be more promising than the linear one.



(a) ACF

(b) pdf

Figure 6.7: Autocorrelation function (a) and probability density function (b) of the case with $K = 6$, quadratic EMR, and $n\_level = 5$.

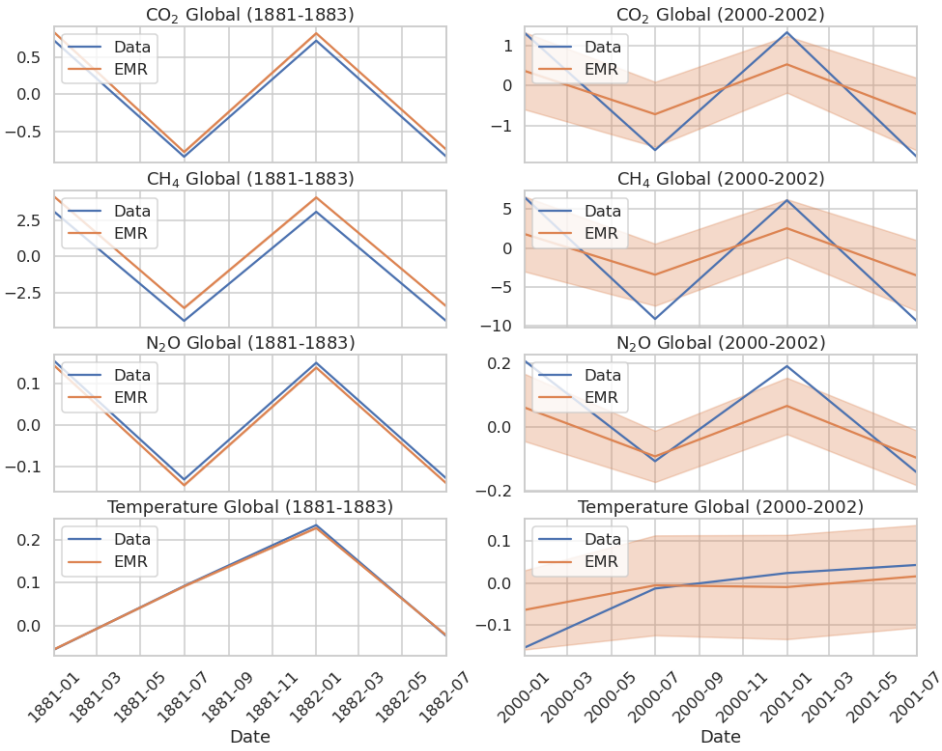


Figure 6.8: Comparison between the original data and the simulation of the case with $K = 6$, quadratic EMR, and $n\_level = 5$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

6.3 Alternative versions of the original time series

One singular behavior found during the analysis of the datasets in Chapter 5 could be seen in the noise component of the greenhouse gases time series. As shown in Appendix C, the noises of all three components have a particular behavior: before a certain date they show a really low amplitude, as compared to a higher one afterwards. This peculiarity, probably caused by the re-elaboration of the data, may have affected the performance of the EMR model. In order to establish if some solution were necessary, three new datasets have been built from the original one.

The first one is identical to the original one, but only the period 1985-2014 is considered. From the year 1985, in fact, the noise component has approximately a higher amplitude for all the components. A drawback is that it is a really small dataset, with 360 data per component against more than 1600 in the original one.

The second dataset substitutes the original noise with a uniform artificial noise (called "noise of type 1"). It is based on four white noises (one per each component) but with a modification. Each noise is given a standard deviation equal to the ones of the original noise.

The third and last dataset also has an artificial noise (called "noise of type 2"). In this case, the base is still a white noise, but it has been transformed so that its maximum amplitude is equal to the one of the original one. Essentially, a noise with the same amplitude as the higher part of the original one has been generated.

In the next sections, the results obtained from applying the three new datasets will be reported in the same way as done with the original time series. To begin with, it will be reported a general overview of the coefficients R^2. Then, an in-depth focus will be dedicated to some particular cases.

6.3.1 Results with the time series 1985-2014

Table 6.2 presents the results obtained considering only the period between 1985 and 2014 in the EMR algorithm. It is already visible how the algorithm has difficulties in stabilizing the model, with R^2 values much higher and more variable than the ones in the previous case.

Another piece of information is the difference in results regarding the coefficient of sub-sampling K and the equation degree if compared to Table 6.1. For example, quadratic EMR models are supported both with $K = 1$ and $K = 3$. This could have been the result of the shorter time series used, having fewer chances of divergences in a small period of time.

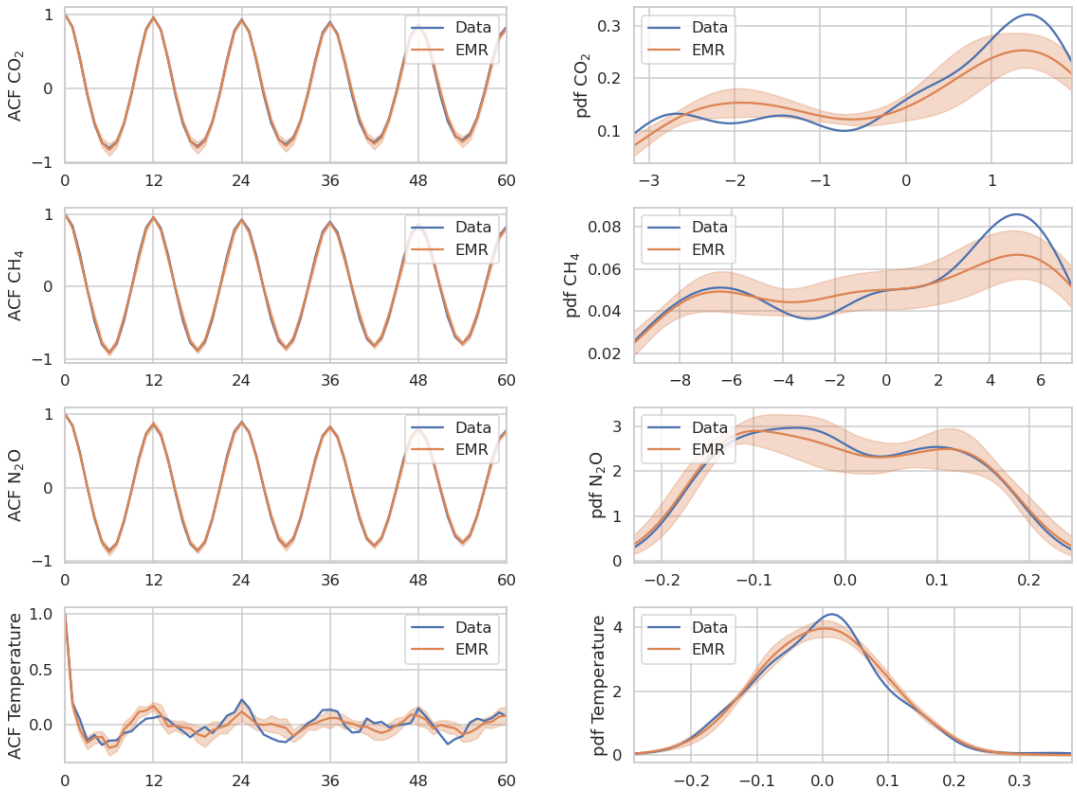
As done with the original dataset, some deeper examinations will be conducted, studying the autocorrelation functions and probability density functions of the simulated data.

	$K = 1$				$K = 3$				$K = 6$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.78	0.21	x	x	0.88	0.17	x	x	0.90	0.16	x	x
2	0.55	0.03	x	x	0.71	0.12	x	x	0.83	0.16	x	x
3	0.70	0.11	0.55	0.03	0.51	0.07	0.53	0.05	0.63	0.06	x	x
4	*0.49*	*0.03*	*0.54*	*0.04*	0.56	0.03	0.56	0.02	0.56	0.03	x	x
5	0.63	0.06	x	x	0.65	0.13	0.56	0.04	*0.54*	*0.04*	x	x
6	0.64	0.09	x	x	0.55	0.06	0.57	0.04	0.59	0.03	x	x
7	0.58	0.06	x	x	0.52	0.05	0.52	0.02	0.60	0.03	x	x
8	**0.51**	**0.01**	x	x	0.54	0.03	0.55	0.01	0.59	0.07	x	x
9	0.52	0.01	x	x	**0.52**	**0.02**	0.56	0.03	0.61	0.07	x	x
10	0.64	0.09	x	x	0.56	0.03	**0.52**	**0.01**	0.65	0.08	x	x

Table 6.2: R^2 results applying the 1985-2014 dataset. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $K = 1$, linear EMR, $n\_level = 8$

This might be the best case obtained using the current dataset.
The autocorrelation (Figure 6.9(a)) seems to be steady over time with the GHGs components. Notwithstanding the temperature anomalies, the dynamics are finely represented. Good results can be found even with the probability density functions (Figure 6.9(a)).
The direct consequence is a system able to nicely simulate the base signal over time, as shown in Figure 6.10.



(a) ACF (b) pdf

Figure 6.9: Autocorrelation function (a) and probability density function (b) of the case with $K = 1$, linear EMR, and $n\_level = 8$.

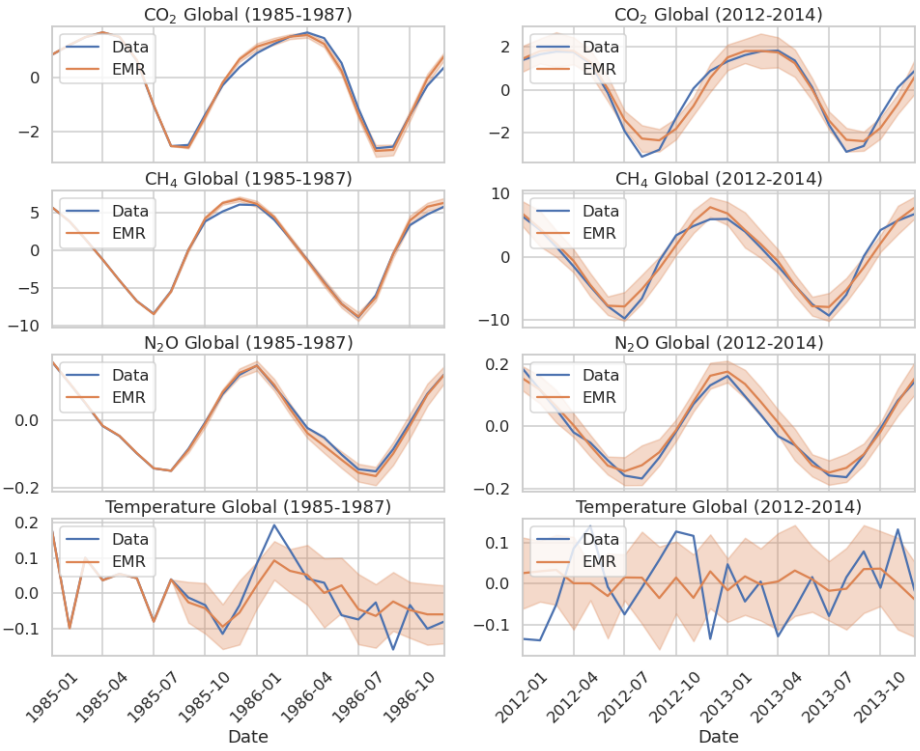
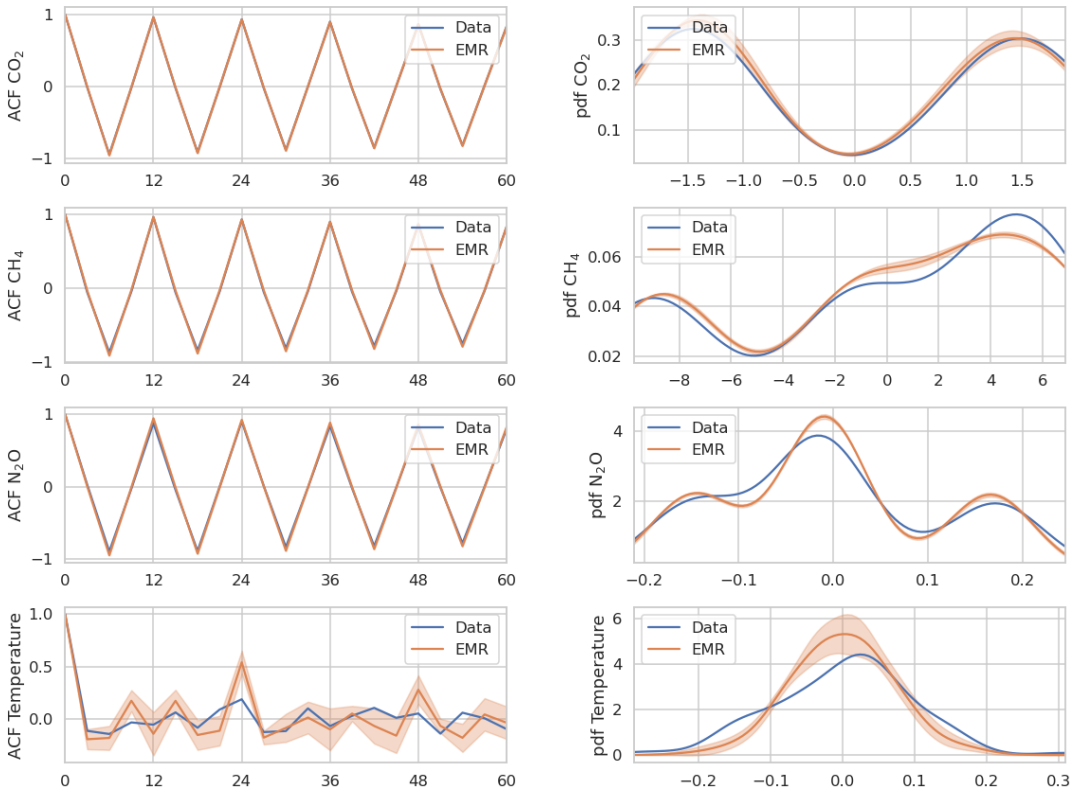


Figure 6.10: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 1$, linear EMR, and $n\_level = 8$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 3$, linear EMR, $n\_level = 9$

By applying a sub-sampling with $K = 3$ the results seem to be good.
The pdf (Figure 6.11(b)) is really precise with low variations. Also, the ACF is able to remain similar to the base one, even for temperature signals.
The resulting simulation keeps really similar dynamics to the original one, except with temperature. On the other hand, the simulation of this last component seems to overlap with the original longer than in other cases.



(a) ACF (b) pdf

Figure 6.11: Autocorrelation function (a) and probability density function (b) of the case with $K = 3$, linear EMR, and $n\_level = 9$.

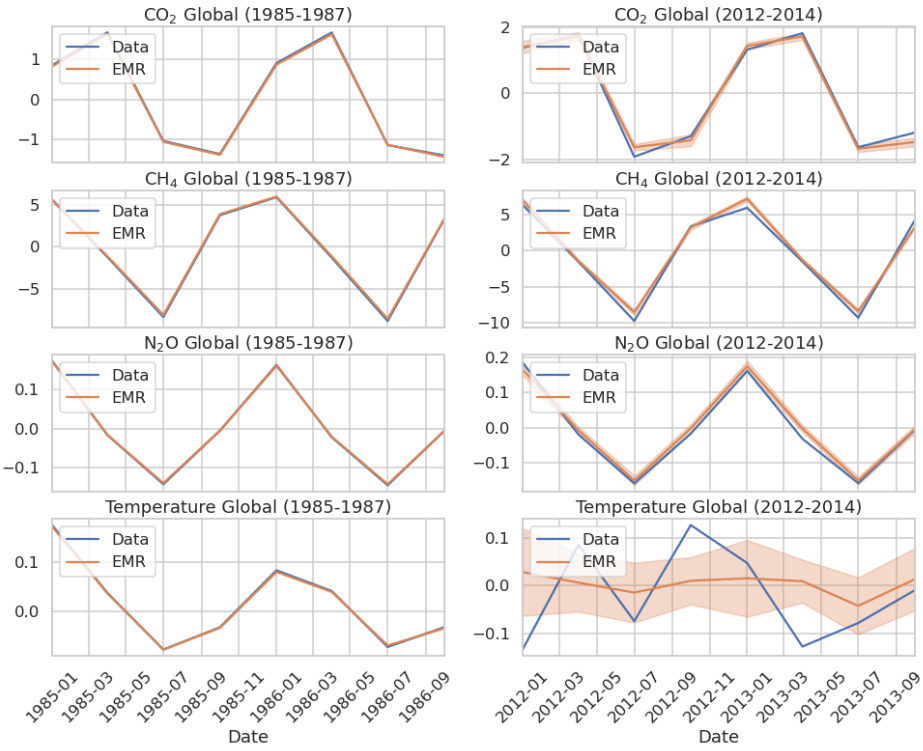
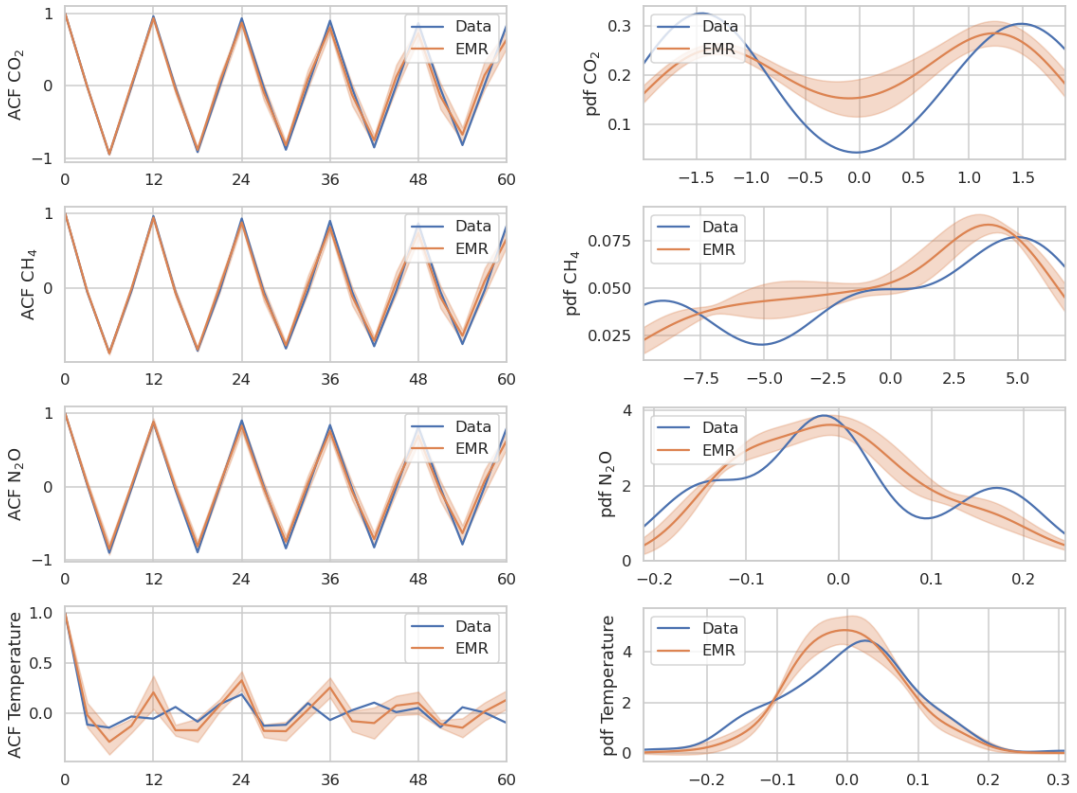


Figure 6.12: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 3$, linear EMR, and $n\_level = 9$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 3$, quadratic EMR, $n\_level = 10$

Unlike the linear EMR model discussed in the last case, the quadratic model seems not to be able to describe nicely the time series.
The autocorrelation function, shown in Figure 6.13(a), seems to deform over time if compared to the original one. The pdfs are also not precise as in the last case. The resulting simulation, although initially precise, deforms over time.



(a) ACF (b) pdf

Figure 6.13: Autocorrelation function (a) and probability density function (b) of the case with $K = 3$, quadratic EMR, and $n\_level = 10$.

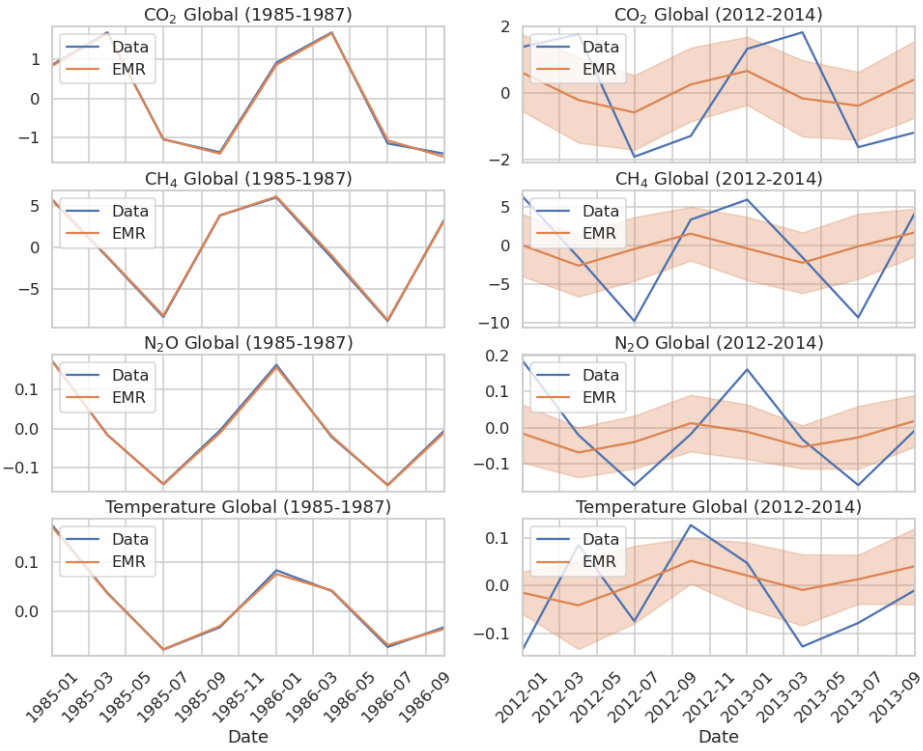


Figure 6.14: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 3$, quadratic EMR, and $n\_level = 10$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

6.3.2 Results with the artificial noise of type 1

The application of an artificial noise of type 1 brought results surely similar to the ones produced using the base dataset. It was impossible to build stable quadratic models without an intensive sub-sampling. It is also interesting to note how, with $K = 1$, the average R^2 coefficients are fairly high. This suggests some struggles of the algorithm into finding an optimal number of levels.

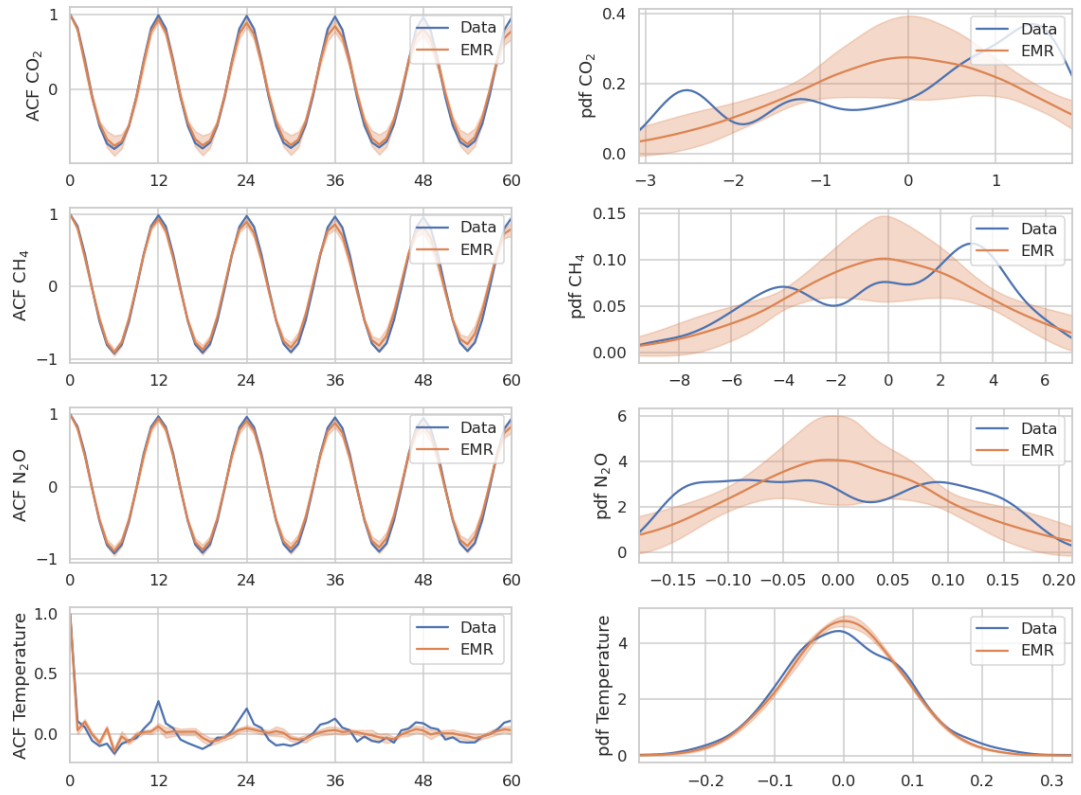
Once again, some particular cases have been selected (in bold and italic in Table 6.3) and will be examined in the following pages.

	$K = 1$				$K = 3$				$K = 6$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.74	0.17	x	x	0.87	0.20	x	x	0.91	0.15	0.92	0.14
2	0.47	0.08	x	x	0.63	0.13	x	x	0.68	0.22	x	x
3	0.67	0.11	x	x	0.72	0.19	x	x	0.50	0.01	*0.51*	*0.02*
4	0.65	0.08	x	x	0.50	0.06	x	x	0.52	0.03	x	x
5	0.57	0.05	x	x	**0.49**	**0.02**	x	x	**0.51**	**0.00**	x	x
6	**0.54**	**0.05**	x	x	0.52	0.02	x	x	x	x	x	x
7	0.55	0.04	x	x	0.52	0.02	x	x	0.51	0.01	x	x
8	0.59	0.09	x	x	0.52	0.02	x	x	x	x	x	x
9	0.61	0.11	x	x	x	x	x	x	x	x	x	x
10	0.56	0.05	x	x	x	x	x	x	x	x	x	x

Table 6.3: R^2 results applying the dataset with artificial noise of type 1. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $K = 1$, linear EMR, $n\_level = 6$

Being this the best case with $K = 1$, it surely shows that the EMR is unable to build linear models using the complete version of this dataset.
While the probability density functions are discrete, Figure 6.15(a) shows autocorrelation functions which decrease over time.
The direct consequence, visible in Figure 6.16 is a model unable to stay precise in the long term.



(a) ACF (b) pdf

Figure 6.15: Autocorrelation function (a) and probability density function (b) of the case with $K = 1$, linear EMR, and $n\_level = 6$.

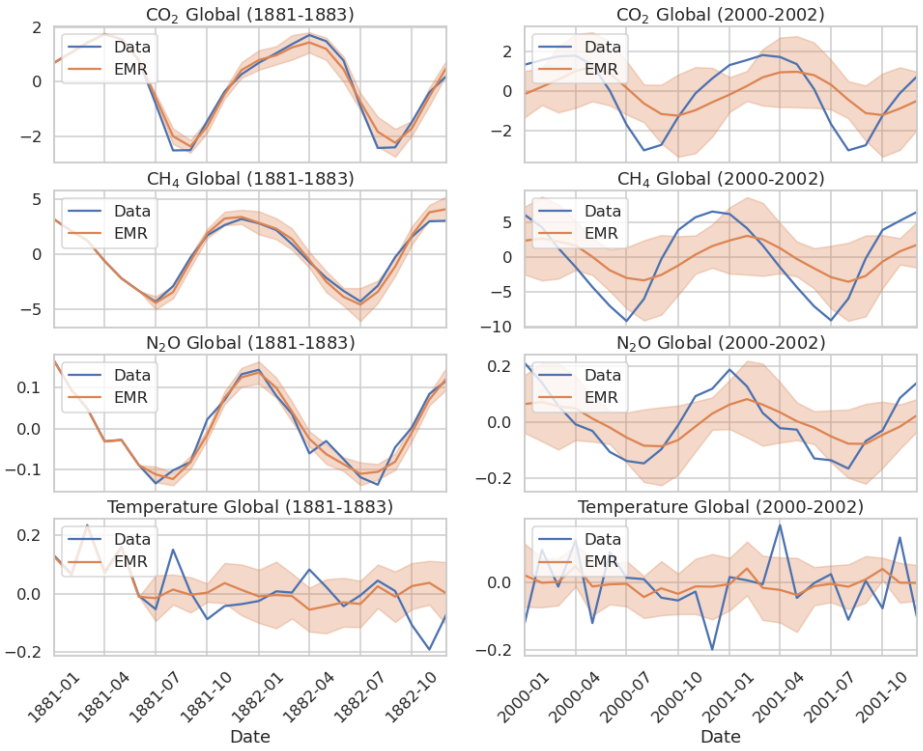
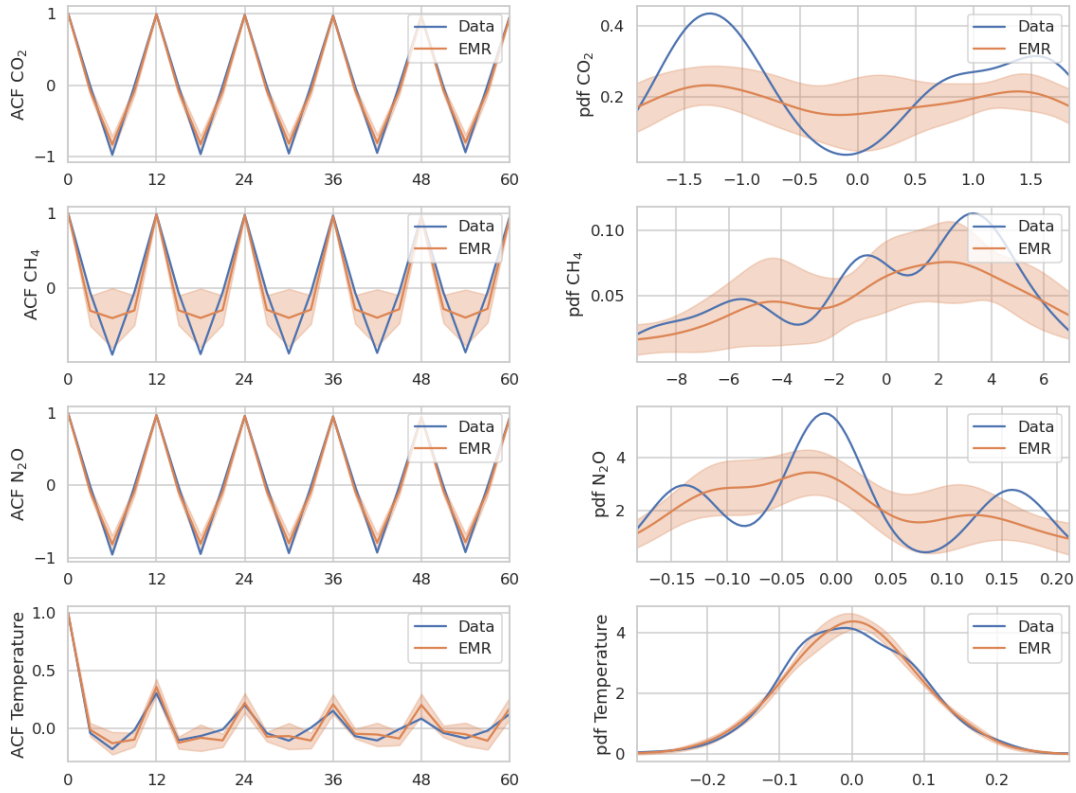


Figure 6.16: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 1$, linear EMR, and $n\_level = 6$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 3$, linear EMR, $n\_level = 5$

Applying a sub-sampling with $K = 3$, it is possible to see some improvements in the signal representation.
The probability density functions are not perfect, and the ACFs show some odd results (like the one with CH_4).
Anyway, the final outcome seems to be a signal finely reproduced over time.



(a) ACF

(b) pdf

Figure 6.17: Autocorrelation function (a) and probability density function (b) of the case with $K = 3$, linear EMR, and $n\_level = 5$.

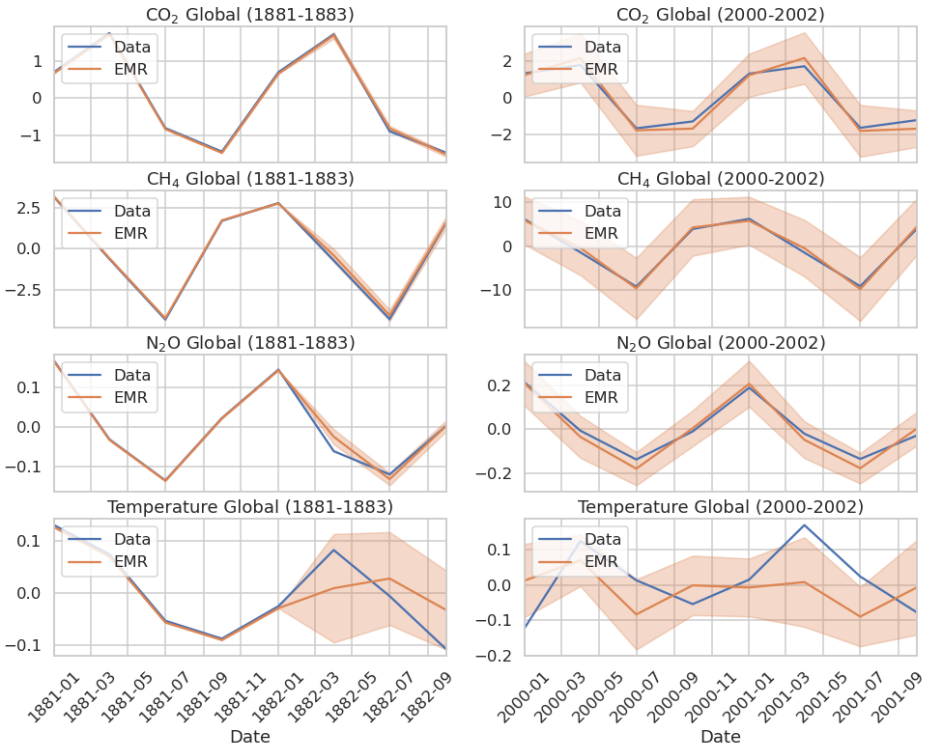
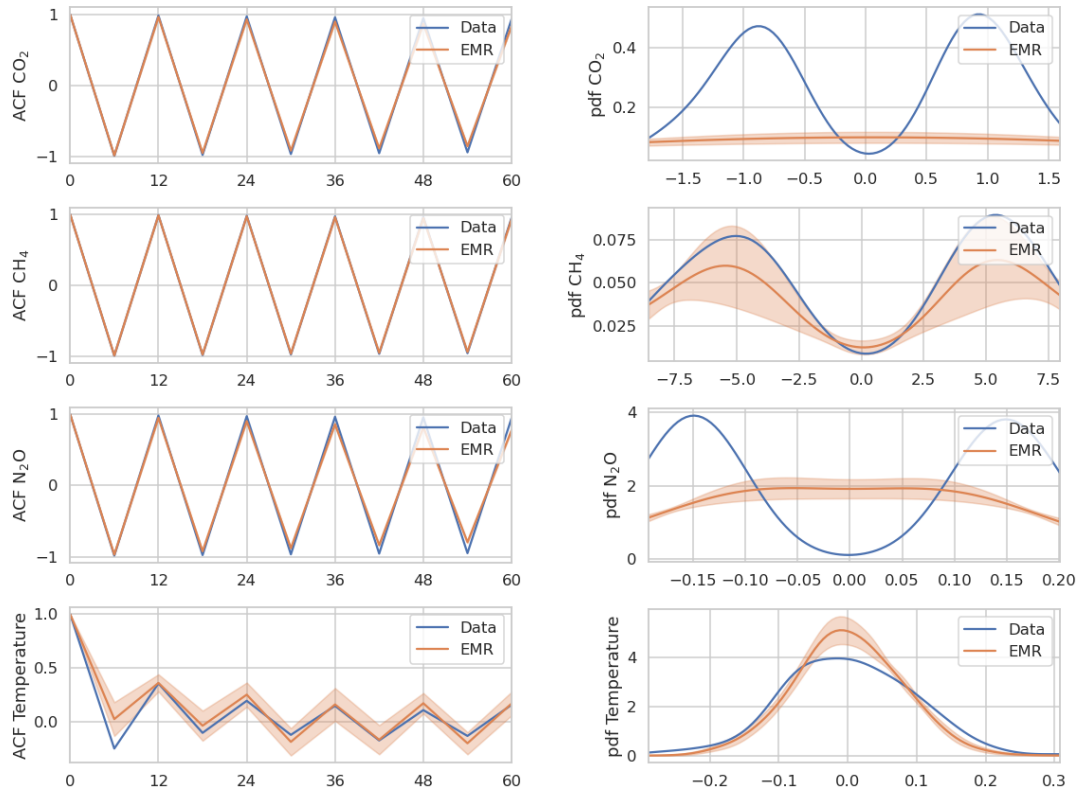


Figure 6.18: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 3$, linear EMR, and $n\_level = 5$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 6$, linear EMR, $n\_level = 5$

Using a sub-sampling even higher than the last case, the results do not seem to improve.
The autocorrelation functions (Figure 6.19(a)) slightly decrease in amplitude over time, and the pdfs are not able to well represent the base signal.
The resulting model shown in Figure 6.20, is unable to well represent data over time.



(a) ACF

(b) pdf

Figure 6.19: Autocorrelation function (a) and probability density function (b) of the case with $K = 6$, linear EMR, and $n\_level = 5$.

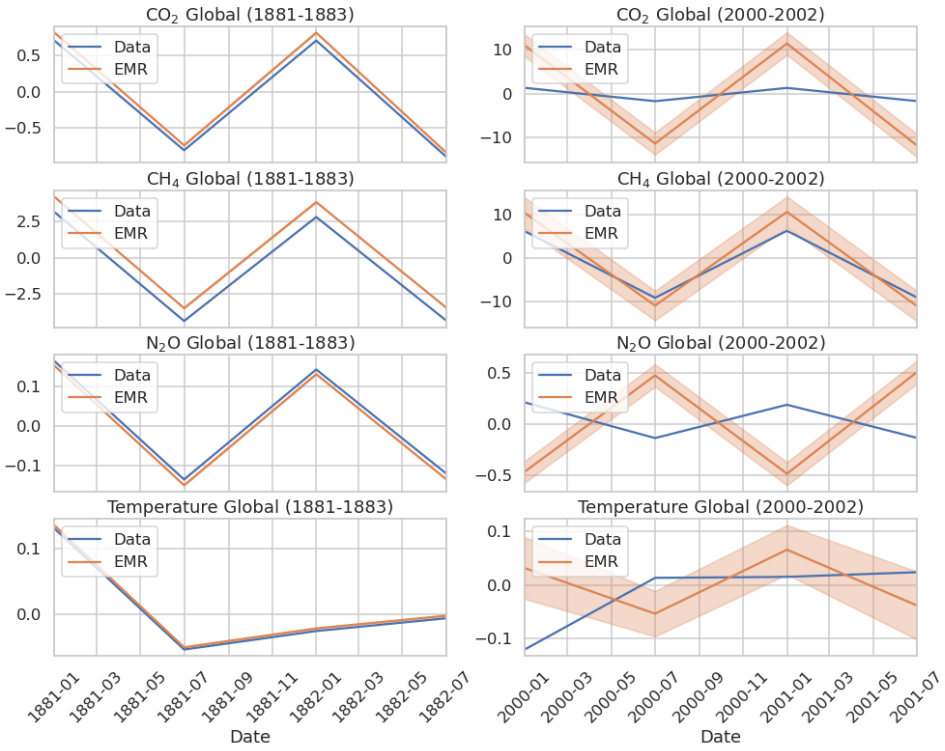


Figure 6.20: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 6$, linear EMR, and $n\_level = 5$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

6.3.3 Results with the artificial noise of type 2

Similar results to the ones obtained with the noise of type 1 have been obtained with noise of type 2 as well. However, one major difference can be found in Table 6.4: with a high number of levels, the algorithm seems able to build quadratic models without the necessity of sampling the time series. Moreover, the statistics seem to reveal promising models of the data. In general, the R^2 mean value seems to be better than the ones in the previous section.

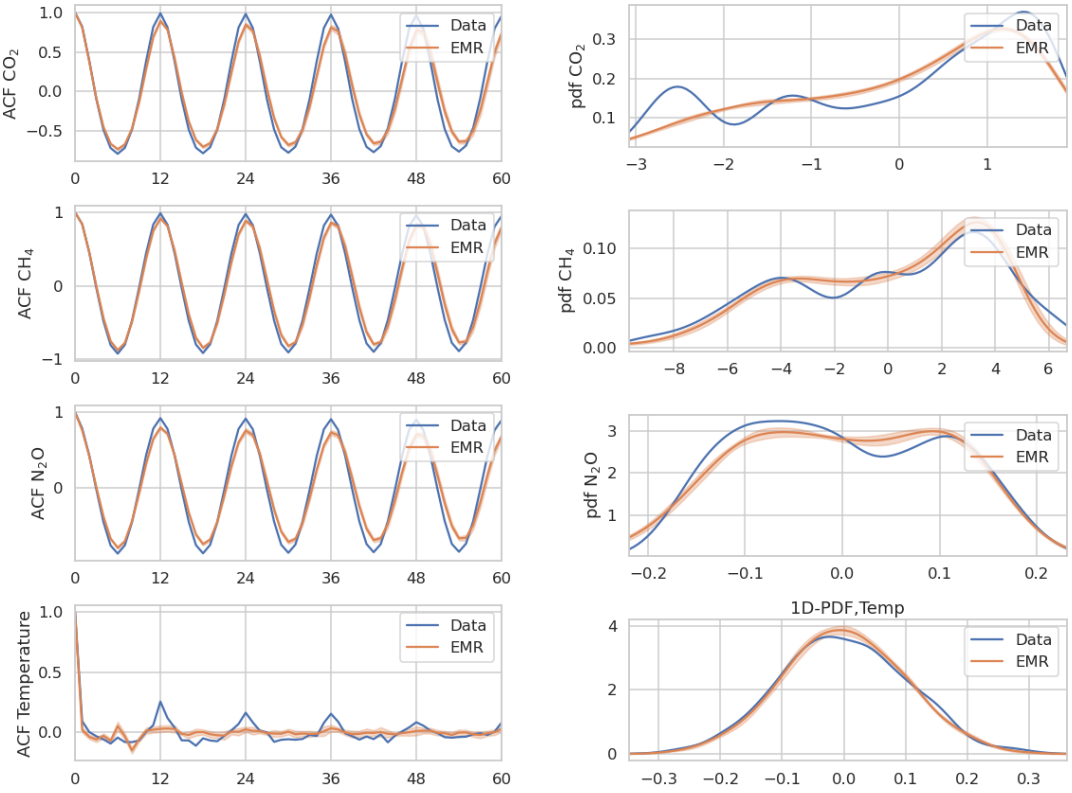
For the last time in this chapter, some relevant cases will be examined, in order to better understand when the algorithm is able to build a well-performing model.

	$K=1$				$K=3$				$K=6$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.72	0.16	x	x	0.85	0.20	x	x	0.89	0.17	0.90	0.16
2	0.45	0.09	x	x	0.62	0.15	x	x	0.68	0.21	0.53	0.07
3	0.63	0.08	x	x	0.71	0.20	x	x	0.50	0.06	0.53	0.02
4	0.62	0.08	x	x	0.49	0.06	x	x	0.52	0.02	**0.50**	**0.01**
5	0.55	0.04	x	x	**0.49**	**0.01**	x	x	0.51	0.01	x	x
6	*0.53*	*0.04*	0.53	0.04	0.52	0.02	x	x	0.51	0.01	x	x
7	0.57	0.07	0.53	0.03	0.52	0.01	x	x	**0.50**	**0.00**	x	x
8	0.60	0.09	0.52	0.01	x	x	x	x	x	x	x	x
9	0.61	0.11	**0.51**	**0.01**	x	x	x	x	x	x	x	x
10	0.55	0.05	0.51	0.01	x	x	x	x	x	x	x	x

Table 6.4: R^2 results applying the dataset with artificial noise of type 2. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $K = 1$, quadratic EMR, $n\_level = 9$

While Figure 6.21(b) shows that the case model makes a great job in simulating the distribution of the data, it is impossible to say the same about the autocorrelation functions. With all three GHGs, the functions reduce their amplitudes and expand their periods over time.
In the end, the model is unable to well represent the original data, as seen in Figure 6.22.
Unfortunately, the theories elaborated while analyzing the R^2 coefficients turned out to be false.



(a) ACF (b) pdf

Figure 6.21: Autocorrelation function (a) and probability density function (b) of the case with $K = 1$, quadratic EMR, and $n\_level = 9$.

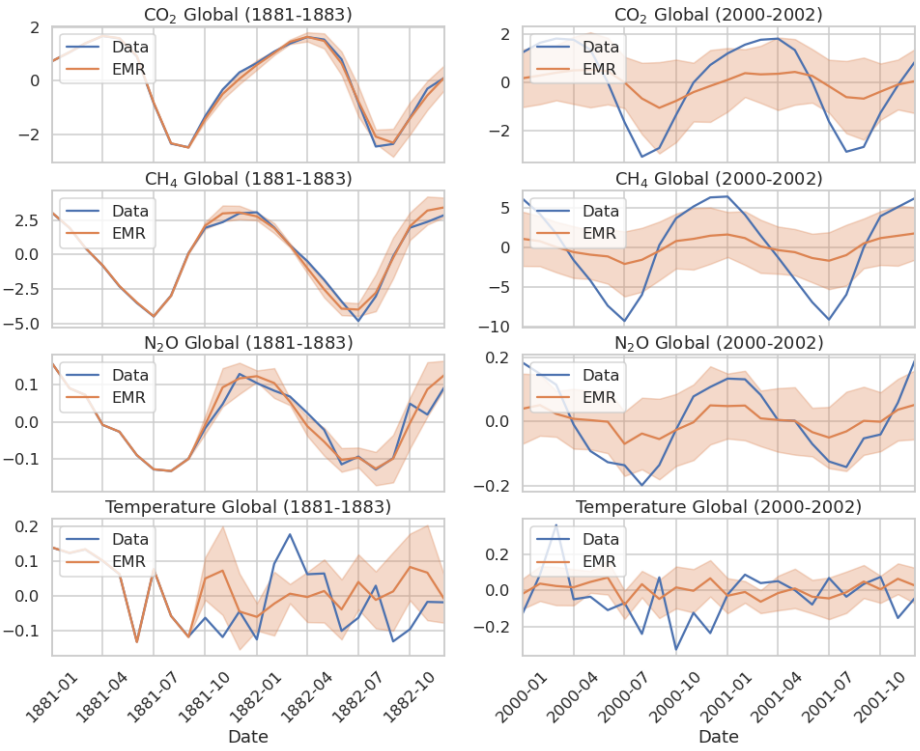
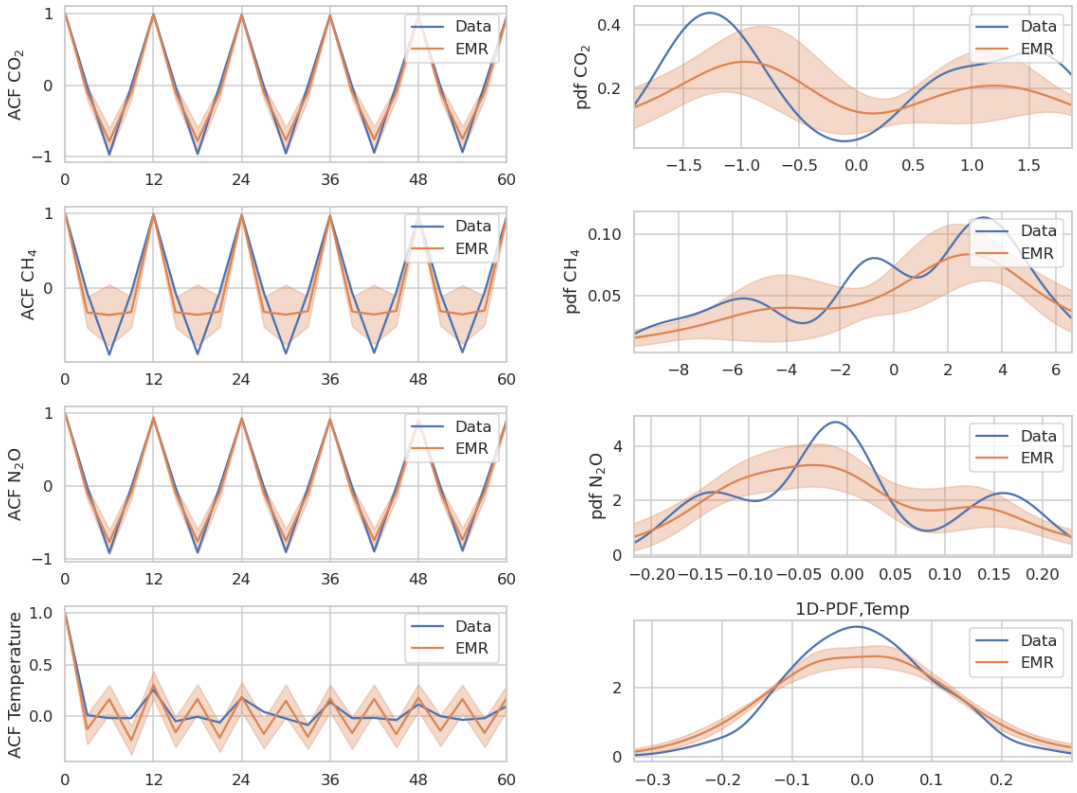


Figure 6.22: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 1$, quadratic EMR, and $n\_level = 9$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 3$, linear EMR, $n\_level = 5$

As usual, using a lighter sampling it is possible to have some improvements in the models. Using this case as an example, it is possible to see in Figure 6.24 a nice representation of the base data, even if not perfect.
It must be underlined that these results have been generated with ACFs and pdfs not really precise, in particular looking at some of the correlation functions in Figure 6.23(a).



(a) ACF

(b) pdf

Figure 6.23: Autocorrelation function (a) and probability density function (b) of the case with $K = 3$, linear EMR, and $n\_level = 5$.

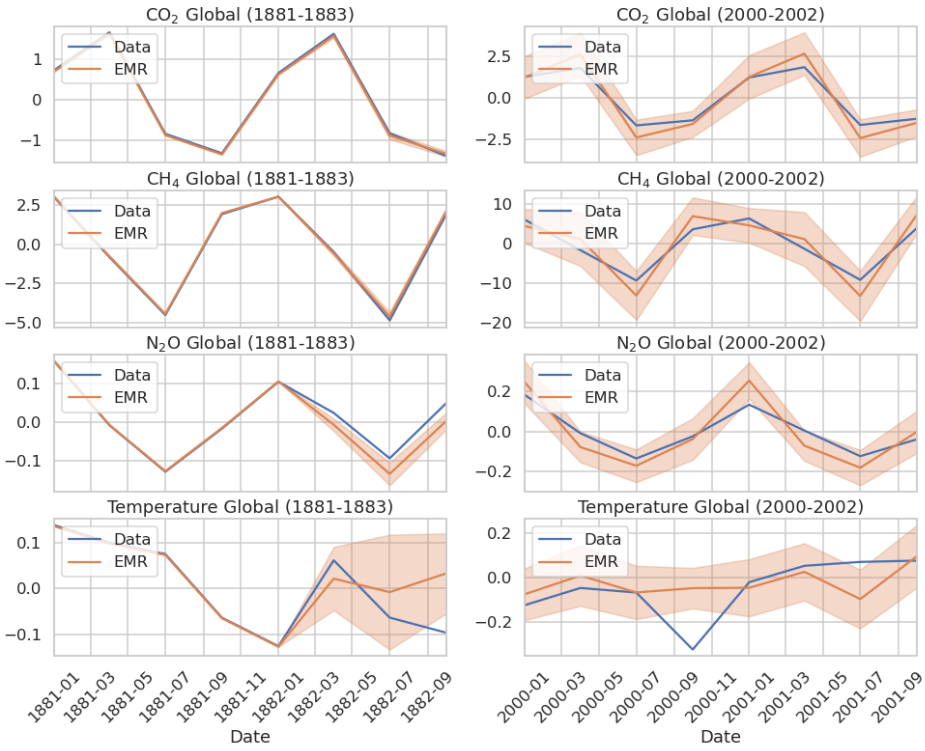
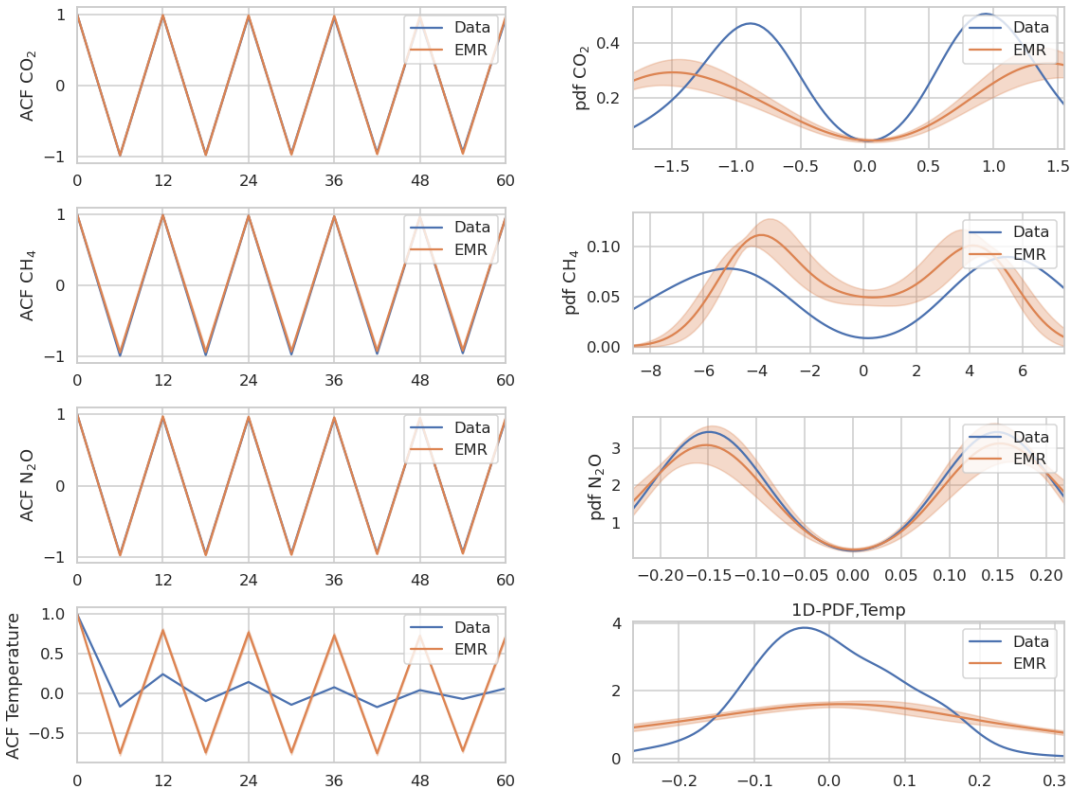


Figure 6.24: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 3$, linear EMR, and $n\_level = 5$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 6$, linear EMR, $n\_level = 7$

While having some precise ACFs of the simulations, the main problems of this case seem to be regarding the probability density function. In particular, the model appears to be completely unable to describe the temperature anomalies, at least from what is possible to observe from the functions.
Figure 6.26 confirms the theory, with temperature data completely different from the base ones. An unexpected outcome is that the model seems to have difficulties also representing methane data.



(a) ACF (b) pdf

Figure 6.25: Autocorrelation function (a) and probability density function (b) of the case with $K = 6$, linear EMR, and $n\_level = 7$.

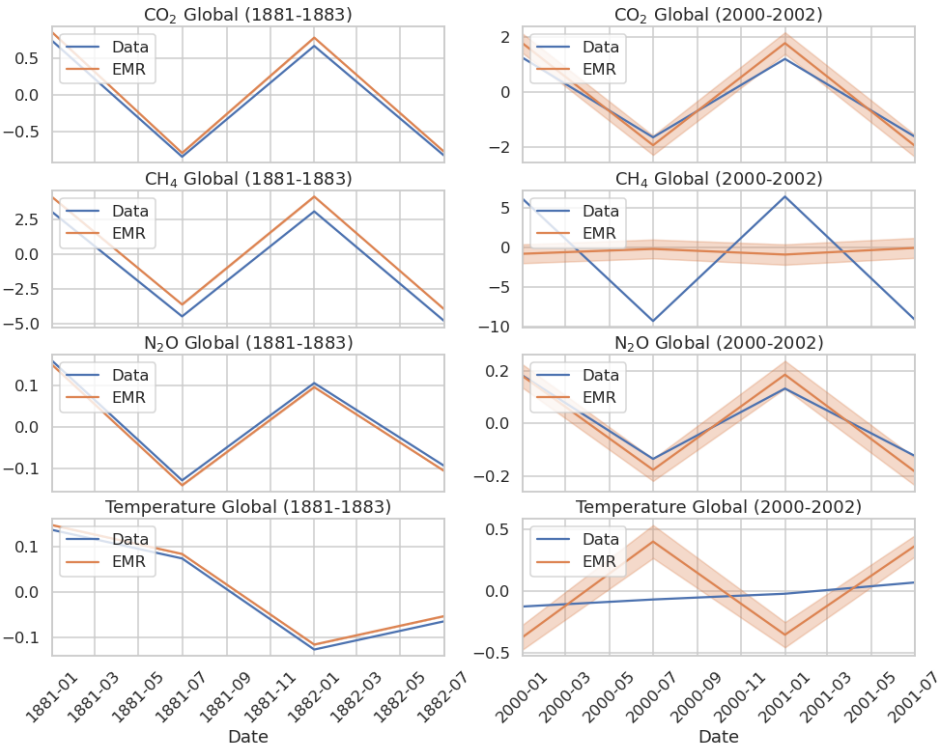
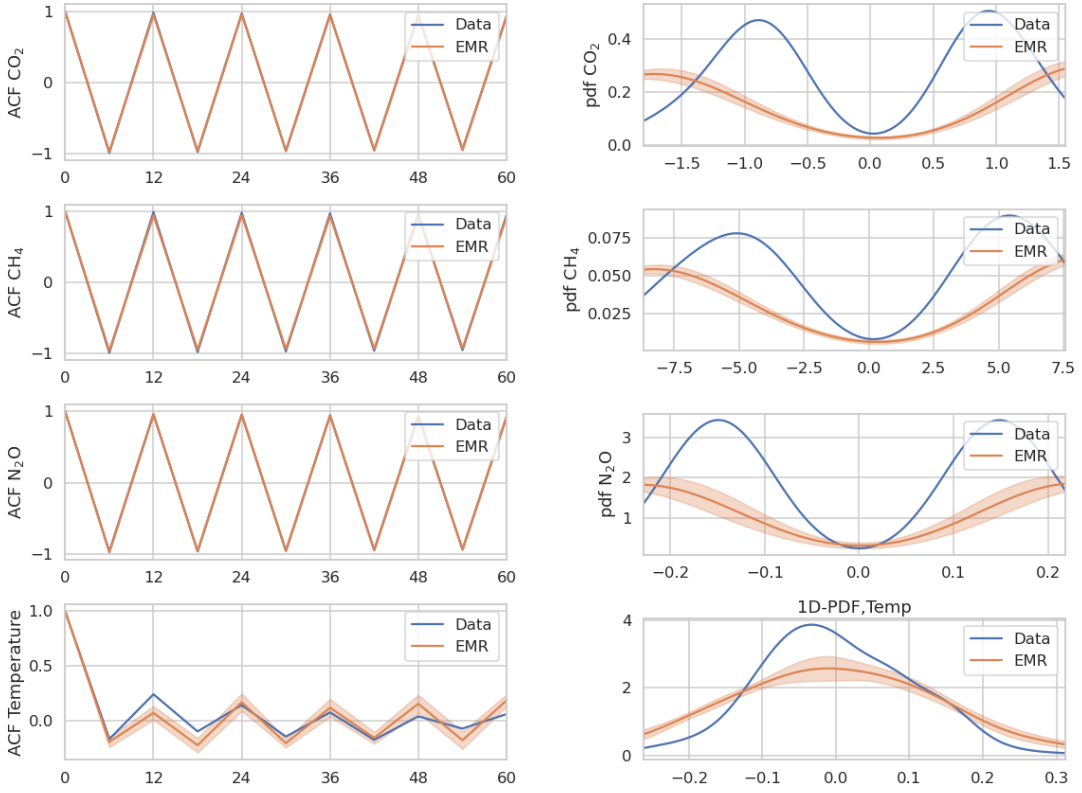


Figure 6.26: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 6$, linear EMR, and $n\_level = 7$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

Case $K = 6$, quadratic EMR, $n\_level = 4$

In the case of $K = 6$, the quadratic model seems to bring far better results than the linear one.
While not having optimal probability density functions, Figure 6.27(a) shows a great capacity in the representation of the autocorrelations, even in the case of temperature anomalies.
By studying directly the data in Figure 6.28, it is possible to see that the model is able to nicely simulate the dynamics over time, in particular regarding the ones relative to the CO_2 and CH_4 components.



(a) ACF

(b) pdf

Figure 6.27: Autocorrelation function (a) and probability density function (b) of the case with $K = 6$, quadratic EMR, and $n\_level = 4$.

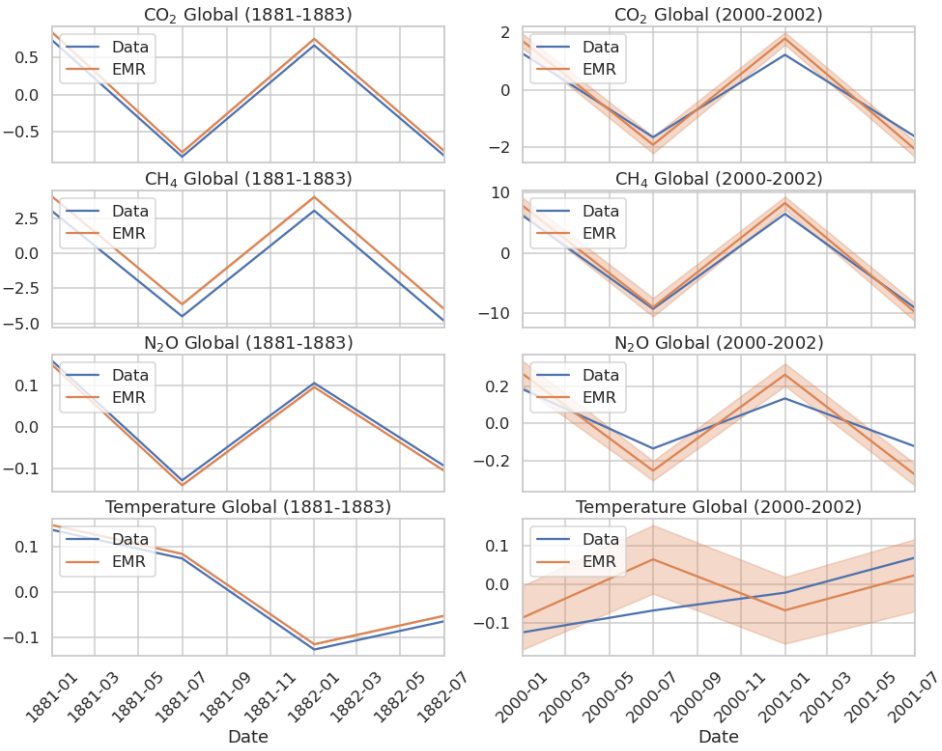


Figure 6.28: Comparison between the 1985-2014 dataset and the simulation of the case with $K = 6$, quadratic EMR, and $n\_level = 4$. The focus periods are 1985-1987 (left) and 2012-2014 (right).

6.4 Final remarks

Based on the current results, the empirical model reduction methodology seems to be suitable to reproduce the dynamics of the climate data which were used. Having analyzed several cases using different datasets and different combinations of hyperparameters, it is possible to report some initial observations.

To begin with, the algorithm seems to be more able in modeling data with a marked seasonality. In many cases, the simulation of the greenhouse gas concentrations is more stable for a longer time. On the other hand, the temperature anomalies seem more difficult to study: often the autocorrelation functions overlap the original ones for only the first year, and this can be confirmed also observing the data themselves. Due to the difficulties in reproducing it nicely, along with the fact that often the pdf of the temperature is well represented as a gaussian function with mean value equal to 0, it is conceivable that the algorithm classifies the temperature data as a noise signal.

While on the subject of ACFs and pdfs, it is important to note how fine representations of the autocorrelations are more important than a nice estimation of the values in the data. In all the cases with a precise simulation of the original data, was more common to have a precise autocorrelation function than a probability density function. In general, the ACFs of the greenhouse gases have to remain constant and don't decrease faster than necessary. Otherwise, the model signal amplitude might become lower and lower up to be equal to 0.

Another important observation is the relationship between the parameters K and n_level. With lower values of K (so if the sampling is less intensive) a higher quantity of levels is necessary in order to obtain the best results. It might be correlated to the length of the time series: longer signals (consequences of lower values of K) are more difficult to model, and a higher quantity of residual levels might be useful for the purpose.

Finally, the best models were obtained with K equal to 1 and 3. Using climate data, it is easy to see why. With $K = 1$ the time series consists of one measure for each month, while with $K = 3$ they consist of four measures for each year: January, April, July, and October. These four months are in the middle of the four seasonal trimesters, so they are well suited to describe the seasons' cycle.

6.4.1 Results with alternative versions

By applying variations in the signals used it is possible to better understand the inner dynamics of the algorithm and the data themselves.

The resulting models perform generally better than the complete time series when considering only the period 1985-2014. One of the reasons might probably be that it is easier to build a model which needs to simulate a shorter signal.

Moving forward, the application of an artificial noise surely had particular effects on the construction of models, like the possibility to build quadratic models with $K = 1$. However, the results showed performances generally far lower than the ones on the original dataset. The main hypothesis is that what was supposed to be only a random noise component in the signal was carrying important information for the system dynamics. It is possible that the replacement of the noise with an artificial one might have removed these dynamics, thus bringing difficulties in the modeling.

Chapter 7

Linear Regression applied to Empirical Model Reduction

The core step in the empirical model reduction method is the calculation of the model coefficients, in order to describe the time series evolution using the equations(EMR). In the Algorithm 1, this step is executed by resolving, for each level l, the matrix equation:

$$\boldsymbol{dx}[l] = \boldsymbol{A}[l] \times \boldsymbol{x_A}[l],$$
$$\boldsymbol{A}[l] = \boldsymbol{dx}[l] \times \boldsymbol{x_A^{-1}}[l],\ 0 \leq l \leq L-1$$

Of course, this is the most direct and suitable solution, but other options can be taken into consideration.

In this chapter, some of these solutions will be evaluated. Like in Chapter 6, multiple models obtained with different combinations of hyperparameters will be calculated. Then, some particular cases will be further examined, analyzing the simulated data themselves, their autocorrelation function, and the probability density function compared to the original ones.

7.1 Linear regression

Linear regression [30] is a quite simple machine learning approach. It assumes the presence of a linear relationship between a predictor variable X and a quantitative response Y. This linear relationship, in its simplest form, can be written as:

$$Y \approx \beta_0 + \beta_1 X$$

Given m observation pairs $(x_1, y_1), \ldots, (x_m, y_m)$ the objective of linear regression is to calculate estimations $\hat{\beta}_0$ and $\hat{\beta}_1$ of the model coefficients such that

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i,\ i \in \{1, \ldots, m\}$$

In order to determine the best coefficients, it is necessary to calculate how close the model is in predicting the observations. One of the most common way is through the ordinary least-squares estimation, which defines the best coefficients as the ones which minimizes the residual sum of squares between the true observations and the estimations:

$$J(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The function above is called cost function.

In the case of multiple predictors variables, the process is called multiple linear regression. Since in the EMR algorithm, the model is simplified with the matrix multiplication

$$\boldsymbol{f_l} = \boldsymbol{A}[l] \times \boldsymbol{x_A}[l],\ 0 \leq l \leq L-1$$

it is possible to apply the linear regression approach in order to calculate the model coefficients. To do this in the Python environment, it has been used the library `scikit-learn`, in particular its functions relative to linear models.

7.1.1 Results using linear regression

Table 7.1 reports the results obtained using linear regression for the calculation of the coefficients. What is easy to observe is that they are, in terms of R^2 coefficient, almost identical to the original ones (Table 6.1). This shows that the level construction was essentially the same, with some particular exceptions. The first one is with $K = 1$, linear EMR, and $n\_level = 2$, case already examined in the previous chapter. Even having the same standard deviation, the case shows to have a lower mean R^2, closer to the optimal value 0.5. The other change is with the quadratic EMR and $K = 6$: the model with 5 levels (which was the best case with these hyperparameters using the original method) is no longer stable. Instead, the new best case is with 3 levels, previously unstable.

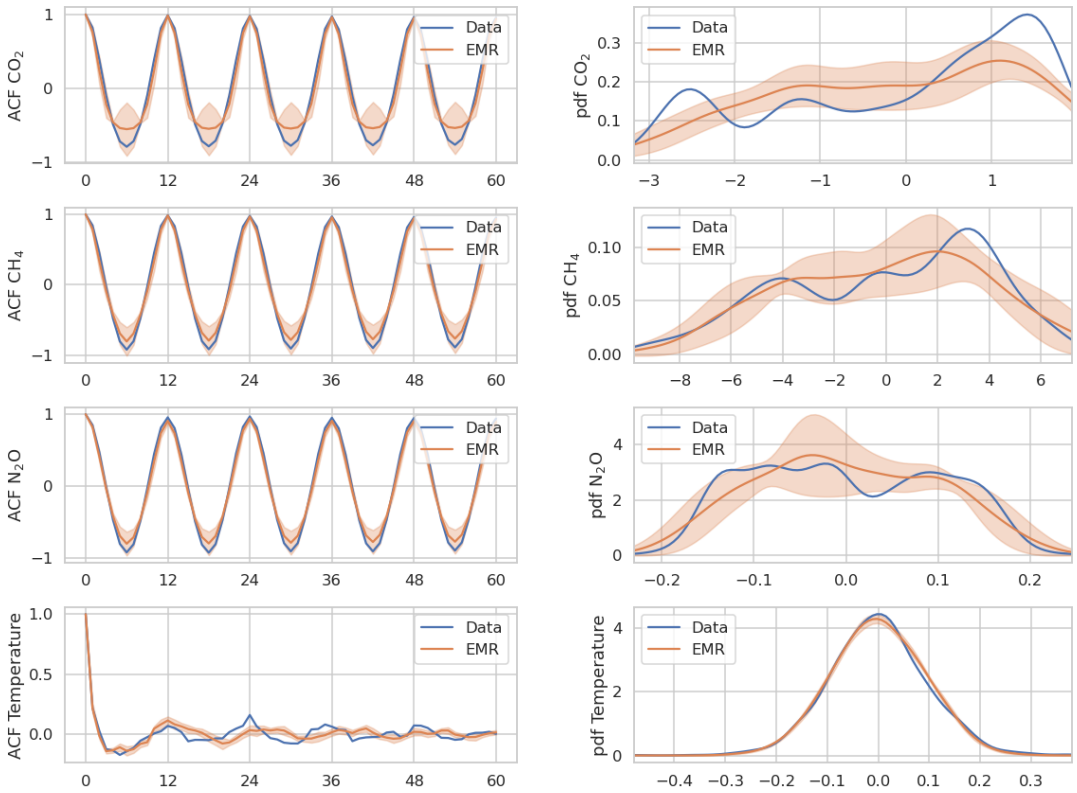
As in the previous chapter, some particular cases will be closely examined, this time also commenting on their performances with respect to their counterparts using the original algorithm.

	$K=1$				$K=3$				$K=6$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.74	0.20	x	x	0.88	0.18	x	x	0.90	0.16	0.91	0.15
2	0.49	0.10	x	x	0.63	0.14	x	x	0.72	0.17	0.54	0.04
3	0.73	0.13	x	x	0.65	0.10	x	x	0.56	0.10	**0.51**	**0.02**
4	0.56	0.04	x	x	0.57	0.07	x	x	*0.53*	*0.08*	0.54	0.02
5	0.58	0.04	x	x	0.63	0.08	x	x	x	x	x	x
6	0.70	0.12	x	x	**0.50**	**0.08**	x	x	x	x	x	x
7	0.53	0.03	x	x	0.49	0.04	x	x	x	x	x	x
8	0.52	0.02	x	x	0.52	0.02	x	x	x	x	x	x
9	**0.52**	**0.01**	x	x	0.54	0.04	x	x	x	x	x	x
10	0.61	0.06	x	x	x	x	x	x	x	x	x	x

Table 7.1: R^2 results using linear regression. For each combination of hyper-parameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $K = 1$, linear EMR, $n\_level = 9$

As per the original algorithm, this was the best case with a linear EMR and $K = 1$. The probability density function seems to be slightly better than the original counterpart, having a value distribution more similar to the base data. However, Figure 7.1(a) shows that the autocorrelation functions are far worse, with distant mean values and higher standard deviations.
The differences in the ACFs are visible also in the data reconstruction, where the amplitude of the simulation signal decreases over time.



(a) ACF (b) pdf

Figure 7.1: Autocorrelation function (a) and probability density function (b) of the case using linear regression with $K = 1$, linear EMR, and $n\_level = 9$.

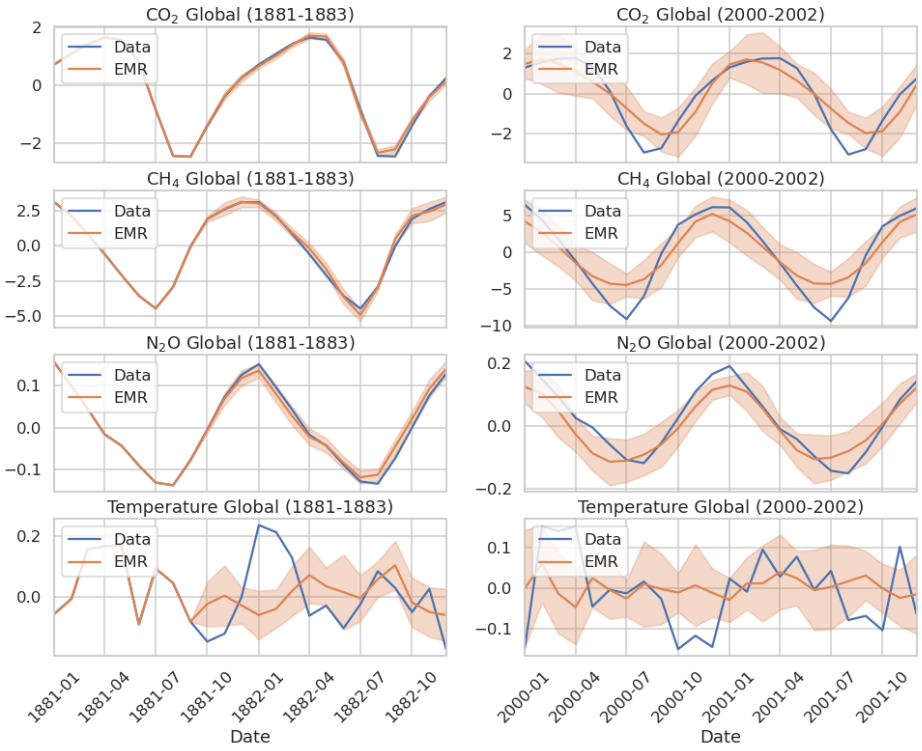
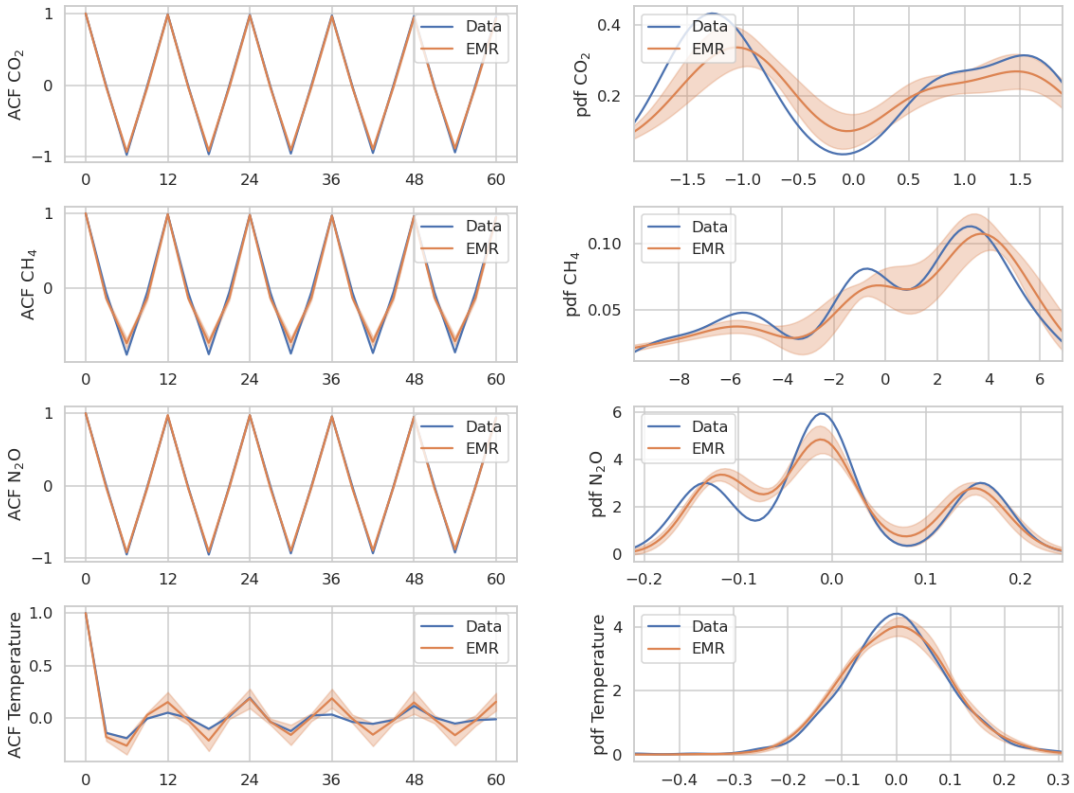


Figure 7.2: Comparison between the original data and the simulation of the case using linear regression with $K = 1$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $K = 3$, linear EMR, $n\_level = 6$

There is not much to say about this case. The functions describing the results are essentially the same as the one in the original algorithm, with probably just slightly better pdfs.
The only relevant aspect might be a better representation of the temperature data, having however also a higher standard deviation.



(a) ACF

(b) pdf

Figure 7.3: Autocorrelation function (a) and probability density function (b) of the case using linear regression with $K = 3$, linear EMR, and $n\_level = 6$.

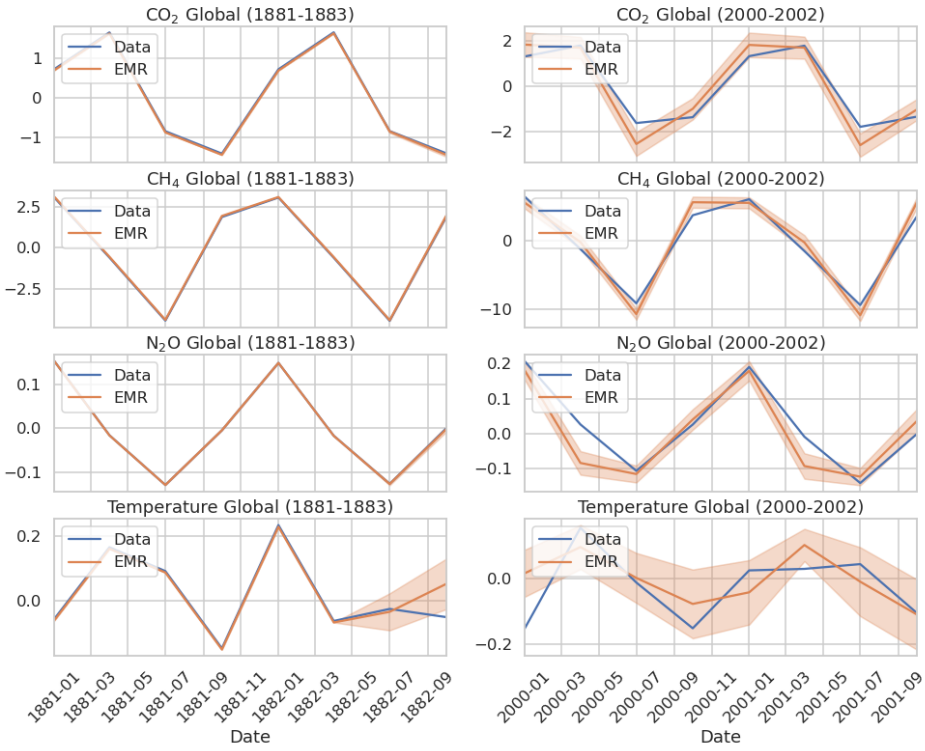
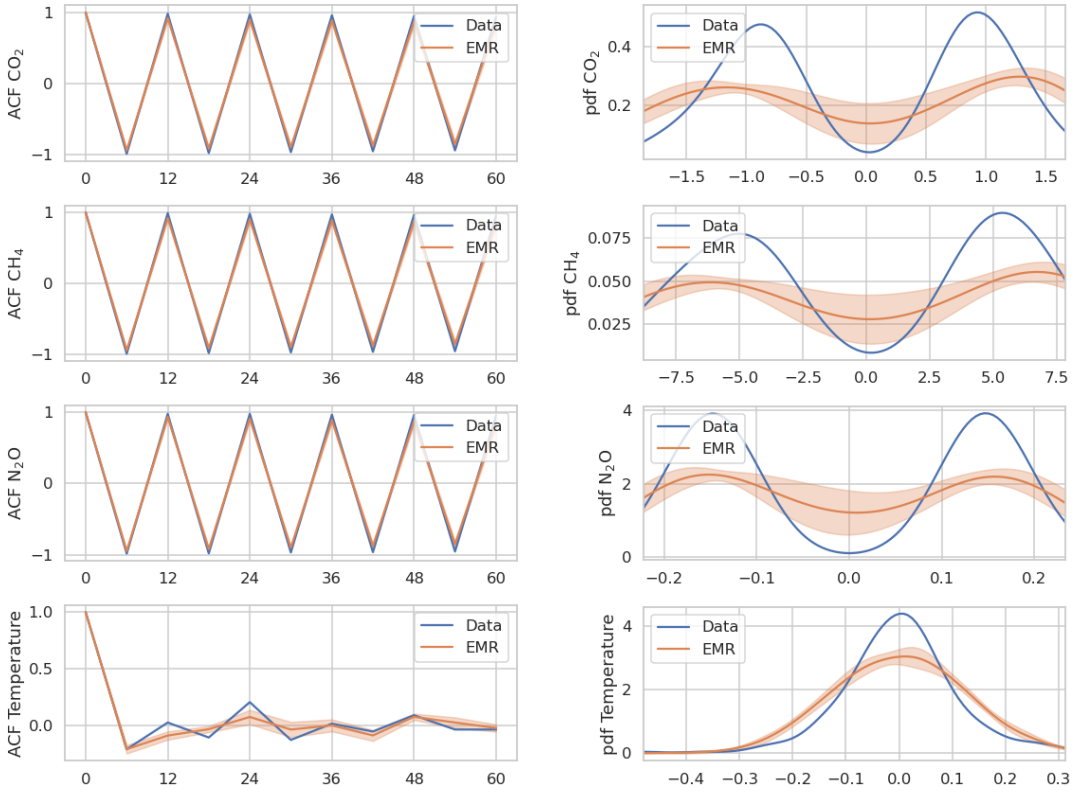


Figure 7.4: Comparison between the original data and the simulation of the case using linear regression with $K = 3$, linear EMR, and $n\_level = 6$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $K = 6$, quadratic EMR, $n\_level = 3$

The results of this model have shown to be much worse than the original ones. Figure 7.5(a) shows that the autocorrelation function of the GHGs seems to be a bit better, with a slower reduction of the amplitude of the function. However, in the case of temperature anomalies, the representation is much worse.
In the end, the simulated signal is not able to reproduce the base data in all the three components, with the amplitudes of the signal that become almost linear over time.



(a) ACF (b) pdf

Figure 7.5: Autocorrelation function (a) and probability density function (b) of the case using linear regression with $K = 6$, quadratic EMR, and $n\_level = 3$.

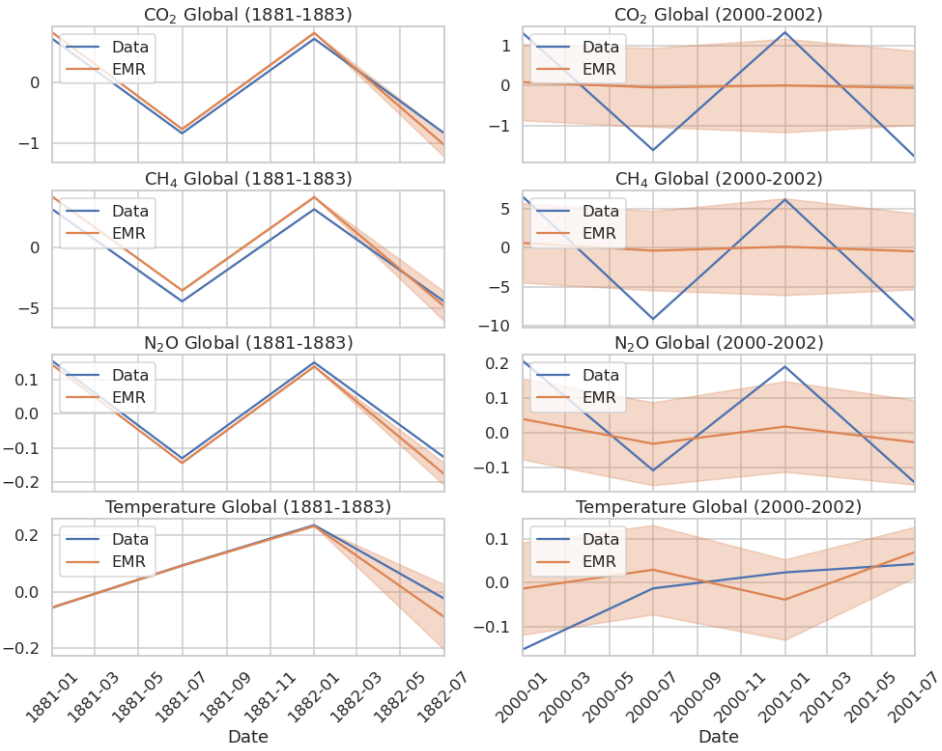


Figure 7.6: Comparison between the original data and the simulation of the case using linear regression with $K = 6$, quadratic EMR, and $n\_level = 3$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

7.2 Ridge and lasso regressions

One of the main issues which can happen during a model training is overfitting: in this case, the model built performs too well with the training observations and brings poor results with any test data used. In order to mitigate this problem, a process called regularization is used. Through a reduction of the magnitude of the model coefficients, it permits to prevent overfitting and reduces the model complexity. Even in the case of an EMR model, which on paper is supposed to fit the original data in the best possible way, it might be interesting to apply some regularization techniques in the model building, thus bringing a better understanding of the dynamics of the empirical model reduction method. A more stable model might help to better forecast future values in the simulations.

In this section, two types of linear regression process with regularization will be used: ridge regression and lasso regression. Both of them rely on adding a penalty in the cost function, proportional to the model coefficients magnitude.

With a ridge regression [31], a L2 regularization is applied, adding to the cost function a penalty equivalent to the square of the value of the coefficients.

$$J_r(\hat{\beta}) = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=0}^{n} \hat{\beta}_j{}^2$$

On the other hand, with a lasso regression [32], a L1 regularization is used, adding a penalty equivalent to the absolute value of the coefficients.

$$J_l(\hat{\beta}) = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=0}^{n} \left|\hat{\beta}_j\right|$$

The coefficient α, present in both formulas, is called the penalty term and it determines the impact of the penalization in the regression. Multiple values of α will be used in order to determine how the regularization might influence the final results, while the value of K will be kept fixed to 1.

7.2.1 Results using ridge regression

Table 7.2 reports the results obtained using a ridge regression. Comparing them with the results obtained using linear regression, is possible to distinguish them into two categories. With small-medium α values, the R^2 values have shown results that are identical to the previous ones. However, with $\alpha = 100$, which brings high penalization in the regression, there are few changes. Above all, there are few stable quadratic models.

In this section, the same case (linear EMR and $n\_level = 9$) will be examined, varying the value of *alpha* and comparing it to the linear regression case. Also, one of the quadratic cases will be studied.

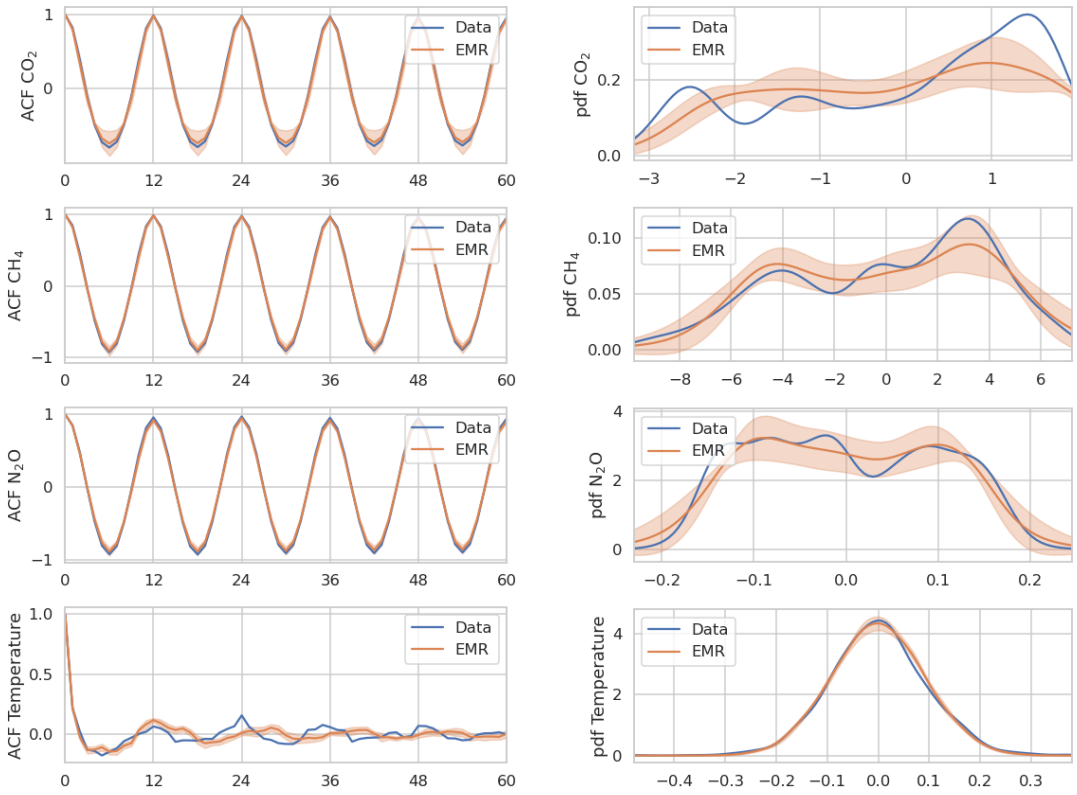
	$\alpha = 0.01$				$\alpha = 1$				$\alpha = 100$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.74	0.20	x	x	0.74	0.20	x	x	0.73	0.20	x	x
2	0.49	0.10	x	x	0.49	0.10	x	x	0.54	0.11	*0.53*	*0.04*
3	0.73	0.13	x	x	0.73	0.13	x	x	0.73	0.13	x	x
4	0.56	0.04	x	x	0.56	0.04	x	x	0.56	0.04	0.54	0.04
5	0.58	0.04	x	x	0.58	0.04	x	x	0.58	0.04	x	x
6	0.70	0.12	x	x	0.70	0.12	x	x	0.70	0.12	x	x
7	0.53	0.03	x	x	0.53	0.03	x	x	0.53	0.03	x	x
8	0.52	0.02	x	x	0.52	0.02	x	x	0.52	0.02	x	x
9	**0.52**	**0.01**	x	x	**0.52**	**0.01**	x	x	**0.52**	**0.01**	x	x
10	0.61	0.06	x	x	0.61	0.06	x	x	0.61	0.06	x	x

Table 7.2: R^2 results using ridge regression. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $\alpha = 0.01$, linear EMR, $n\_level = 9$

This first case already shows how even a small regularization is able to bring results. Overall, both the ACFs and the pdfs are more precise than the ones obtained using linear regression. In particular, the autocorrelation functions (Figure 7.7(a)) are far more stable, with an amplitude more constant in time and lower standard deviations.
Figure 7.8 shows how all these factors bring in the end to more precise simulations.

If the results are compared to their counterparts using the original algorithm (Figures 6.1-6.2) the results are quite similar. Both the ACFs and pdfs of the simulations seem to be better when ridge regression is used. On the other hand, by examining directly the signals, it is possible to see that the cases perform better or worse according to the single component.



(a) ACF (b) pdf

Figure 7.7: Autocorrelation function (a) and probability density function (b) of the case using ridge regression with $\alpha = 0.01$, linear EMR, and $n\_level = 9$.

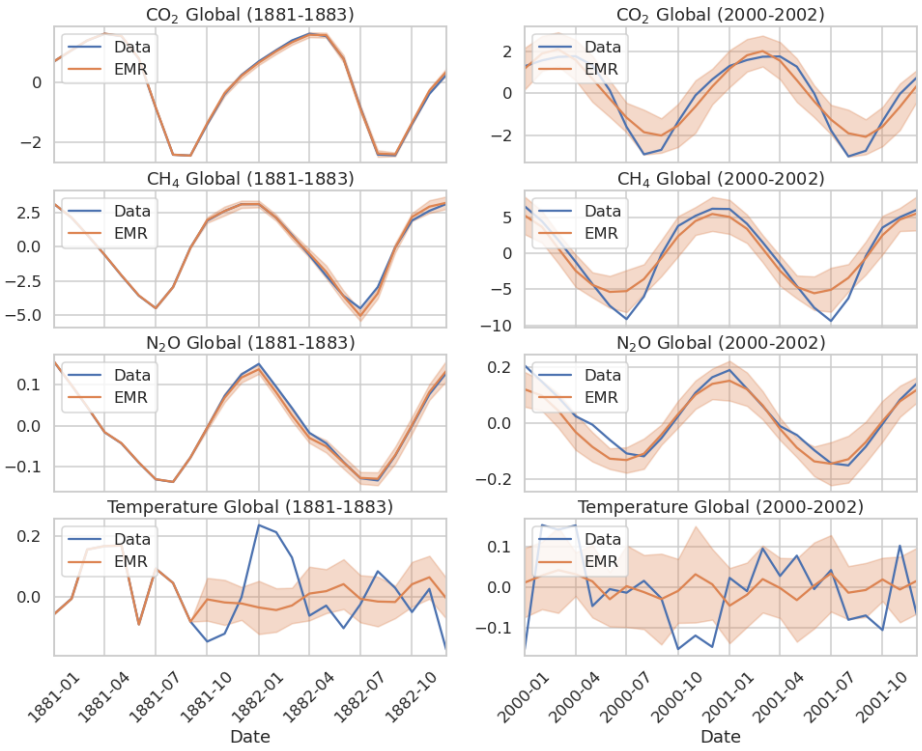
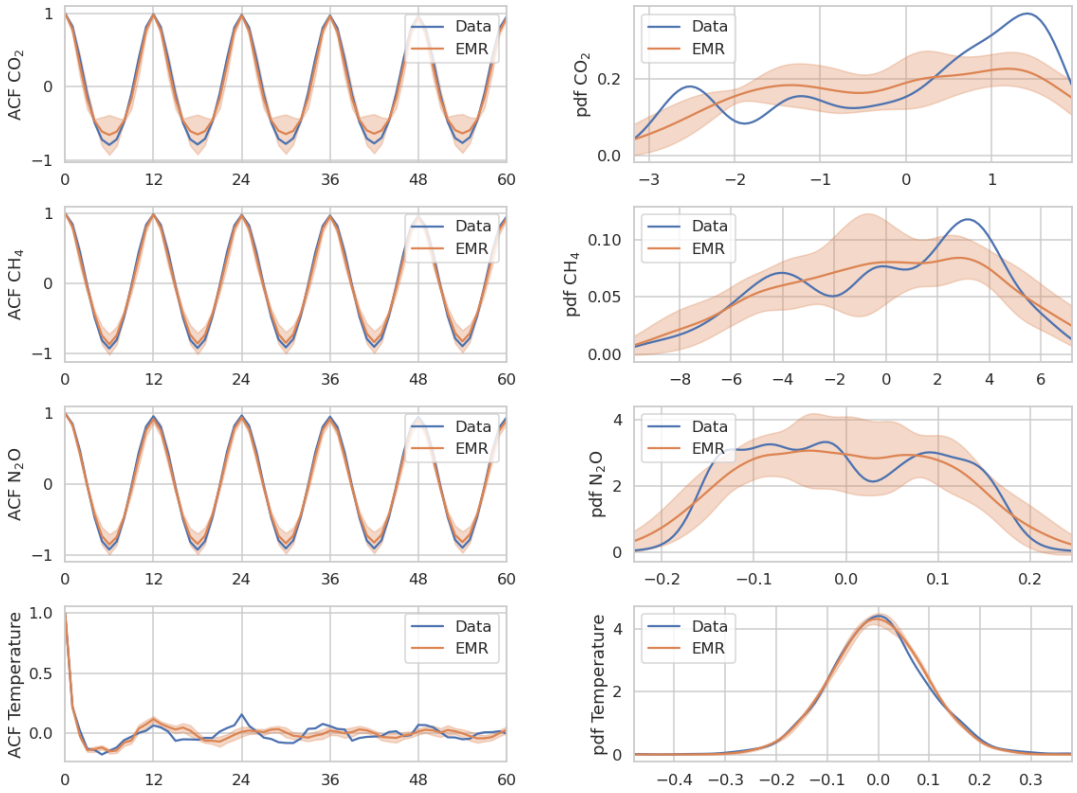


Figure 7.8: Comparison between the original data and the simulation of the case using ridge regression with $\alpha = 0.01$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $\alpha = 1$, linear EMR, $n\_level = 9$

Increasing the penalization, the results seem to get worse even if not as bad as the ones using just linear regression.
Both the ACFs and the pdfs have higher standard deviations. However, the model seems to be able to simulate the data nicely.



(a) ACF (b) pdf

Figure 7.9: Autocorrelation function (a) and probability density function (b) of the case using ridge regression with $\alpha = 1$, linear EMR, and $n\_level = 9$.

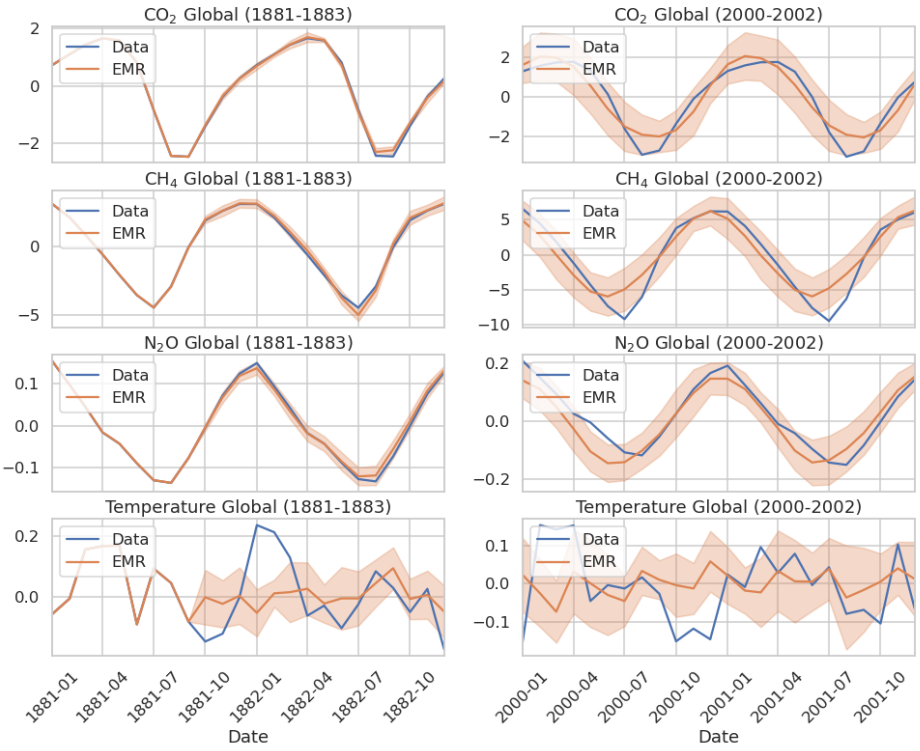
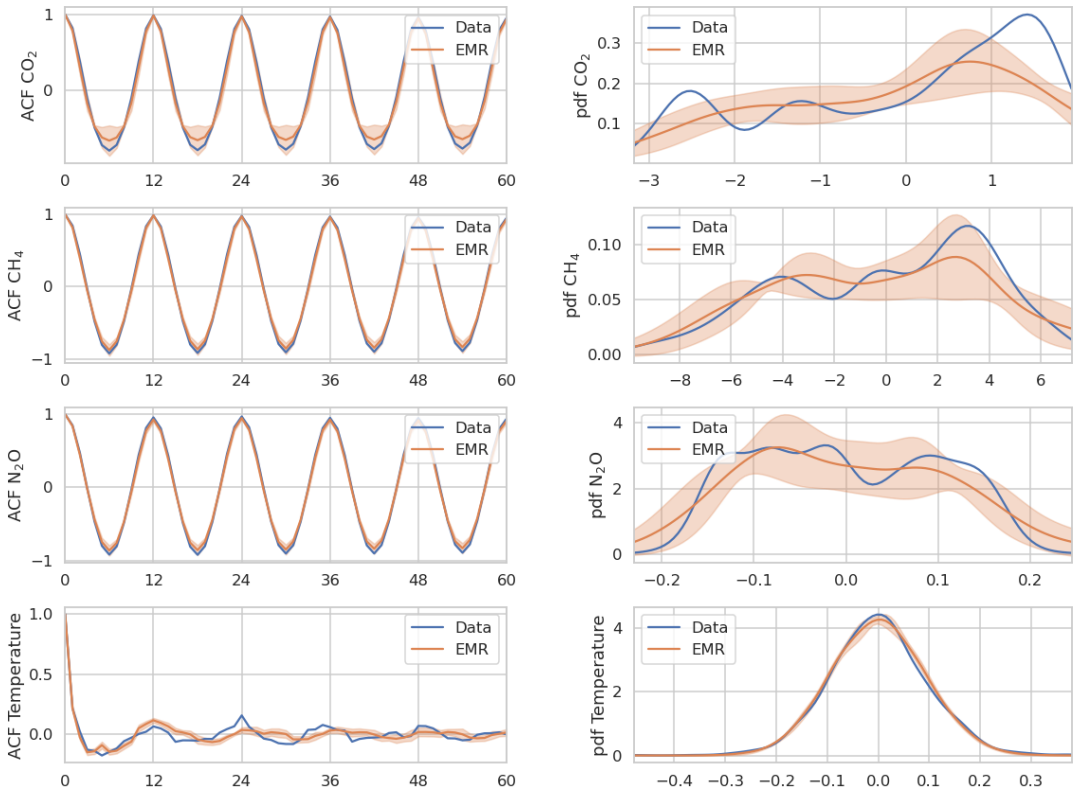


Figure 7.10: Comparison between the original data and the simulation of the case using ridge regression with $\alpha = 1$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $\alpha = 100$, linear EMR, $n\_level = 9$

A further increase of the value of α, which means applying a high penalization, brings results better than with $\alpha = 1$. However, they don't seem to be as good as the ones with $\alpha = 0.01$.
The autocorrelation functions fit well and the pdfs have just standard deviations which are slightly bigger than the ones with $\alpha = 0.01$.
Anyway, the simulations are pretty good, with just a small translation of the periodic cycle which develops over time, as shown in Figure 7.12.



(a) ACF (b) pdf

Figure 7.11: Autocorrelation function (a) and probability density function (b) of the case using ridge regression with $\alpha = 100$, linear EMR, and $n\_level = 9$.

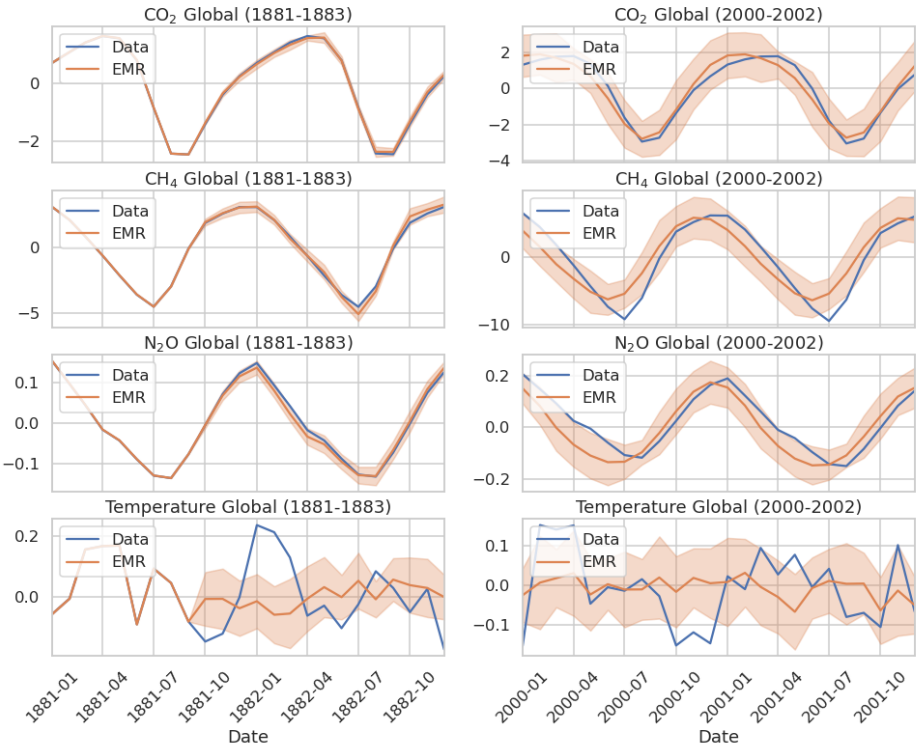


Figure 7.12: Comparison between the original data and the simulation of the case using ridge regression with $\alpha = 100$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

7.2.2 Results using lasso regression

The results of Table 7.3, compared to the ones obtained using ridge regression, are quite singular. While there are minor changes with a small value of α, applying a higher penalization brings an odd behavior: not only the quadratic models are present, but their R^2 coefficients are identical to the ones of the linear models. A possible hypothesis might be that the coefficients relative to the quadratic components of the EMR equations become so small that the model is considered almost identical to a linear one. Anyway, only by examining the cases more deeply, it is possible to establish if there are distinctions between the cases.

In this section, as with the results of ridge regression, the cases with $n\_level = 9$ will be examined and compared to the linear regression ones, this time by varying both the value of α and the degree of the model. The particular case with $\alpha = 0.01$, quadratic EMR, and $n\_level = 4$ will also be considered.

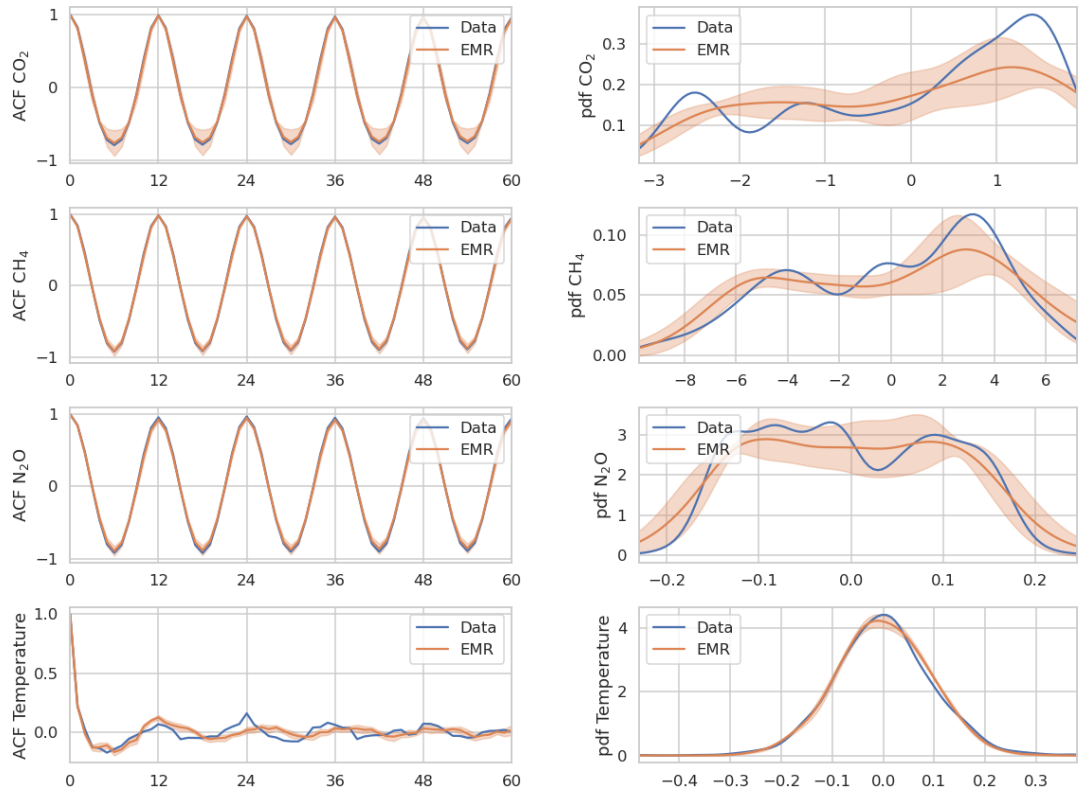
	$\alpha = 0.01$				$\alpha = 1$				$\alpha = 100$			
	linear		*quadratic*		*linear*		*quadratic*		*linear*		*quadratic*	
L	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}	μ_{R^2}	σ_{R^2}
1	0.74	0.20	x	x	x	x	x	x	x	x	x	x
2	0.53	0.10	x	x	0.83	0.04	0.83	0.04	0.83	0.04	0.83	0.04
3	0.73	0.13	x	x	0.73	0.13	0.73	0.13	0.73	0.13	0.73	0.13
4	0.56	0.04	*0.54*	*0.06*	0.56	0.04	0.56	0.04	0.56	0.04	0.56	0.04
5	0.58	0.04	x	x	0.58	0.04	0.58	0.04	0.58	0.04	0.58	0.04
6	0.70	0.12	x	x	0.70	0.12	0.70	0.12	0.70	0.12	0.70	0.12
7	0.53	0.03	x	x	0.53	0.03	0.53	0.03	0.53	0.03	0.53	0.03
8	0.52	0.02	x	x	0.52	0.02	0.52	0.02	0.52	0.02	0.52	0.02
9	**0.52**	**0.01**	x	x	**0.52**	**0.01**	**0.52**	**0.01**	*0.52*	*0.01*	*0.52*	*0.01*
10	0.61	0.06	x	x	0.61	0.06	0.61	0.06	0.61	0.06	0.61	0.06

Table 7.3: R^2 results using lasso regression. For each combination of hyperparameters are reported the mean and standard deviation of the R^2 values over the four components. Results in **bold** indicate cases which are presented in the following pages, the ones in *italic* are other interesting cases.

Case $\alpha = 0.01$, linear EMR, $n\_level = 9$

Once again the results show that the presence of regularization improves the quality of the model developed. The ACFs and the pdfs are far better than the ones obtained with linear regression. This improvement is also visible comparing the simulations.

This model shares similar results to the same case using ridge regression, with precise autocorrelation functions and a nice representation of the data, even when compared to the original model of Figure 6.1(a).



(a) ACF (b) pdf

Figure 7.13: Autocorrelation function (a) and probability density function (b) of the case using lasso regression with $\alpha = 0.01$, linear EMR, and $n\_level = 9$.

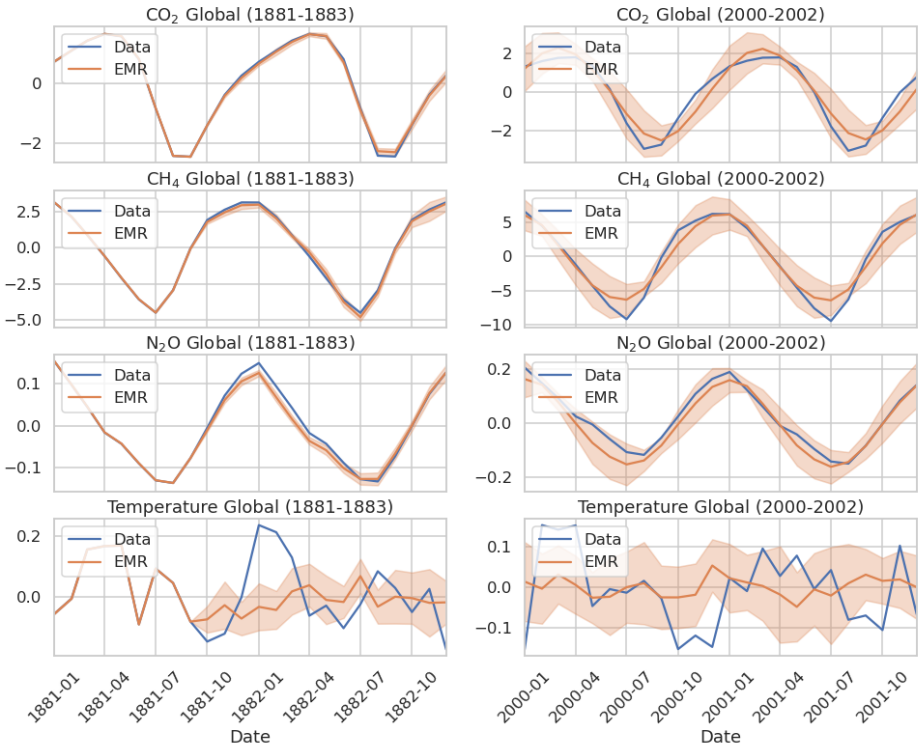
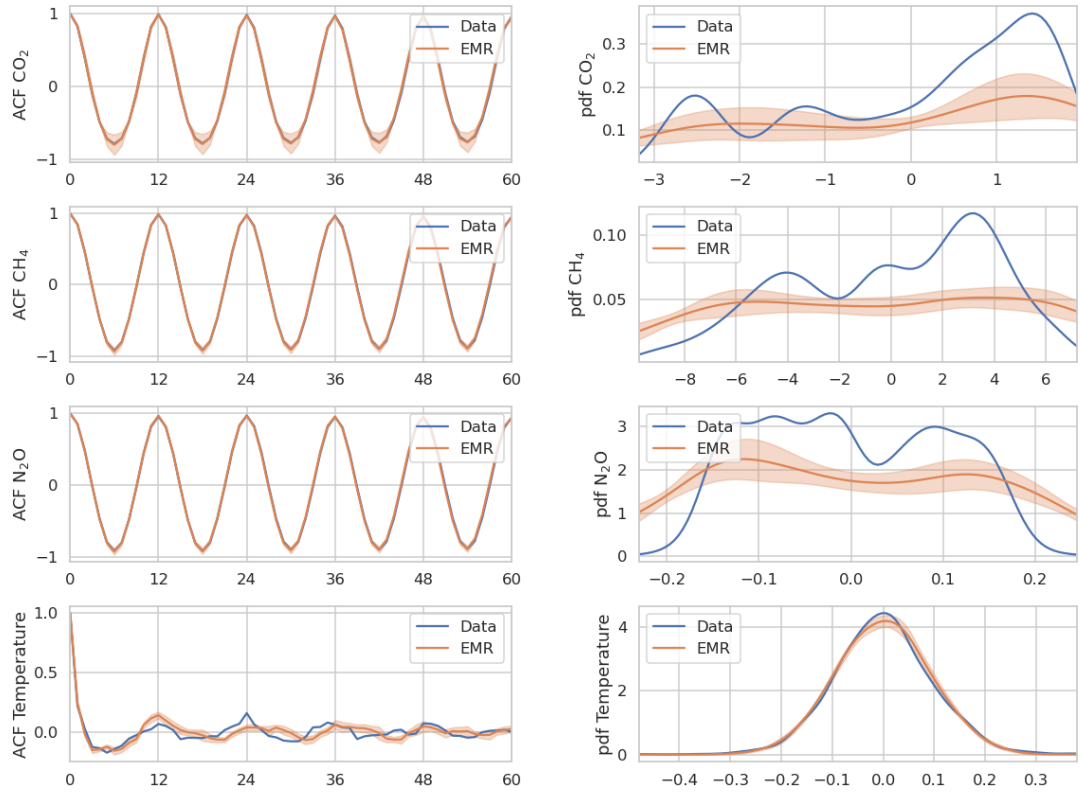


Figure 7.14: Comparison between the original data and the simulation of the case using lasso regression with $\alpha = 0.01$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $\alpha = 1$, linear EMR, $n\_level = 9$

Lasso regression shows to be unable to represent the model using medium or high penalization, as shown in the cases here below. This is the first example.
While the ACFs seem to be really precise, the probability density functions, visible in Figure 7.15(b), showed to be pretty poor in representing the data. The result is a model which is unable to correctly represent the signals.



(a) ACF

(b) pdf

Figure 7.15: Autocorrelation function (a) and probability density function (b) of the case using lasso regression with $\alpha = 1$, linear EMR, and $n\_level = 9$.

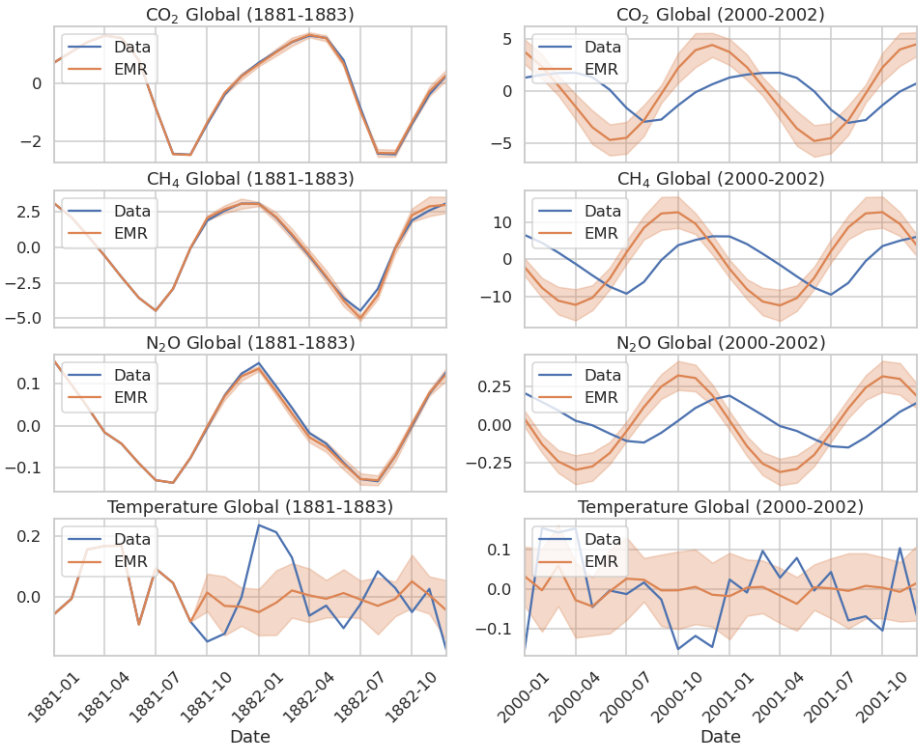
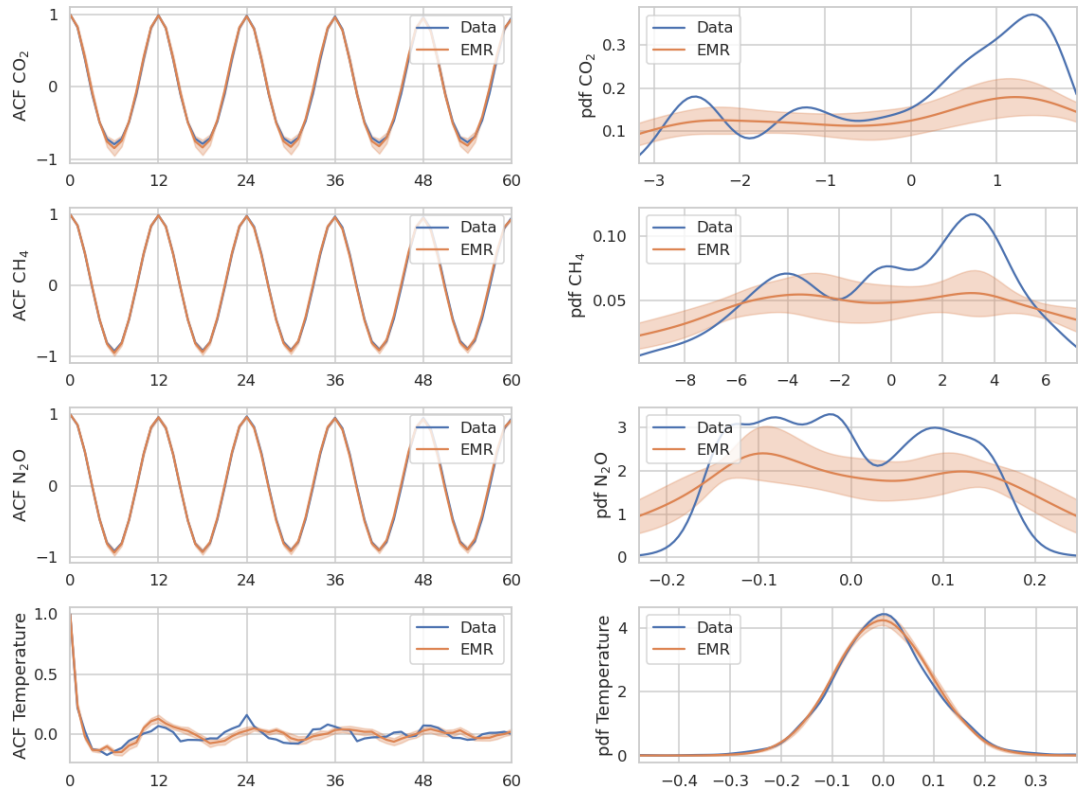


Figure 7.16: Comparison between the original data and the simulation of the case using lasso regression with $\alpha = 1$, linear EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

Case $\alpha = 1$, quadratic EMR, $n\_level = 9$

This case confirms the results obtained as in Table 7.3, specifically that the quadratic models were similar to the linear ones.
Both the ACFs and the pdfs are almost identical to the ones of the linear model. This brings the same results in the simulation, having a poor representation of the data.



(a) ACF

(b) pdf

Figure 7.17: Autocorrelation function (a) and probability density function (b) of the case using lasso regression with $\alpha = 1$, quadratic EMR, and $n\_level = 9$.

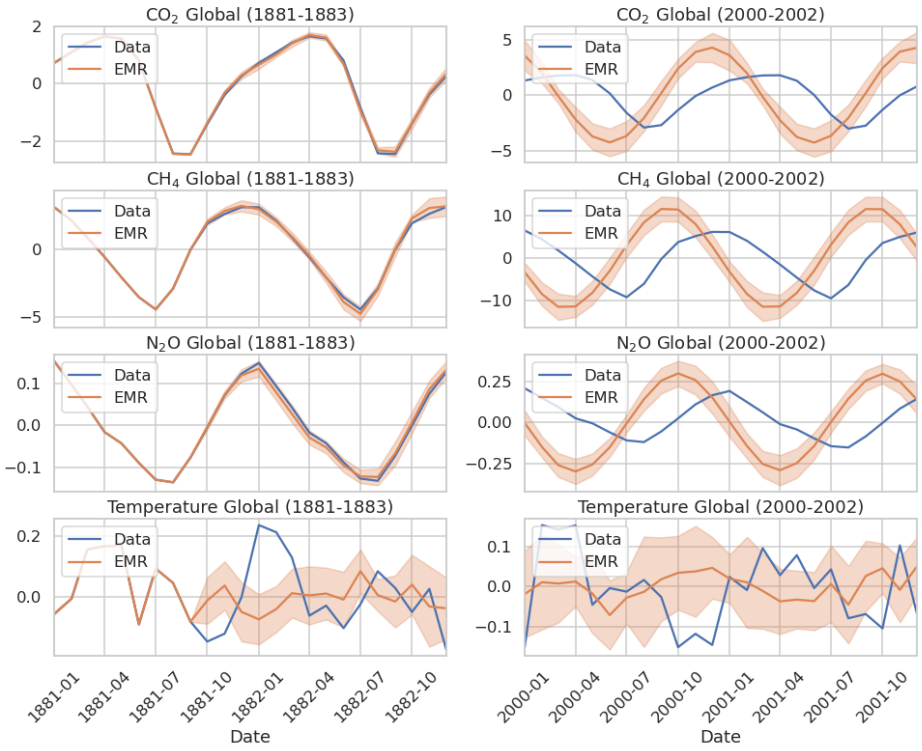


Figure 7.18: Comparison between the original data and the simulation of the case using lasso regression with $\alpha = 1$, quadratic EMR, and $n\_level = 9$. The focus periods are 1881-1883 (left) and 2000-2002 (right).

7.3 Final remarks

The analysis of alternative methods to calculate the model coefficients has surely been fruitful. It was possible to see many different results varying the method applied and the hyperparameters. The multiple linear regression was surely the method with the worst results. In particular, the cases with $K = 1$ were unable to build models as good as the ones of the original algorithm. Only the linear model with $K = 3$ was able to show similar results to its counterpart.

Speaking instead about the application of regularizations, the outcomes were surely more promising. While having almost no impact when choosing the optimal number of levels, it was possible to see different effects in the models themselves.

The models obtained with ridge regression have been shown to obtain really impressive results with both low and high penalizations, but having the best ones in the first case. On the other hand, lasso regression has brought good results, equal to the ones obtained with ridge regression, only applying a small penalization. With values of α higher than 1, the models became all similar and inaccurate.

In the end, ridge regression has proven to be the best method among the three analyzed in this chapter. The models were fairly the more accurate, and the best case ($\alpha = 0.01$, linear EMR, and $n\_level = 9$) could even compete, in terms of performances, with the counterpart built with the original algorithm.

Chapter 8

Conclusions

The objective of this thesis project was to make an exhaustive analysis of the empirical model reduction methodology, and in particular its application on climate data and models. Since it represents an emerging and little discussed argument, there was ample margin for experimentation, combining it with alternative methods and examining the outcomes. Also, its application to climate data permitted to better understand the dynamics involved in general circulation model and earth system models.

An exhaustive analysis of the core concepts of climate research has been reported. In particular, topics like the climate variability, the climate sensitivity and the definition of climate models for simulations were discussed. The empirical model reduction methodology has then been presented, relating the principal theory concepts. The algorithm used has also been described, reporting the principal parameters and the steps executed.

An exhaustive analysis on the data has been performed prior to the experiments with the algorithm. There were four time series involved, three describing the concentrations of the three principal greenhouse gases (CO_2, CH_4, and N_2O) and the last one presenting the temperature anomalies evolution over the years. Both global and hemisphere means data were available, but only the first ones were used for the experiments. The data exploration involved a phase of cleaning and processing, in order to uniform all the data structures, followed by a phase of analysis of the data properties using Fourier transforms, STL decomposition and analysis of the partial autocorrelation functions. The results permitted to reveal a clear distinction between the GHGs (where there was a solid annual seasonality) and the temperature data. This might be due to the fact that the gases data, meant to be used in an earth system model, have been processed before with data assimilation methods. Finally, the four time series were directly compared examining their correlations to each other and revealing a strong correlation between the temperature and the gases concentrations and among the gases themselves.

After this initial analysis, the experiments themselves were conducted with the EMR algorithm, and various results compared as they changed with different combinations of parameters. The same experiments were then performed with modified versions of the dataset, in particular one using recent data and two with artificial noise. Some of the results obtained were surely promising, with models able to faithfully reproduce the signals. In general, linear models revealed to be the more stable and precise than the quadratic ones, indicating a likely presence of linear relations between the four components. However, the quadratic models were able to better represent the data distribution, with probability density functions more similar to the original ones. It was also possible to discover an inverse proportion between the number of levels of the model and the coefficient of sub-sampling. Finally, the experiments executed with the other versions of the dataset brought further interesting results. Above all, the odd behaviours obtained using data with artificial noise lead one to think that the original noise of the data was bringing some fundamental information and that substituting it has brought to the deletion of that information.

The original algorithm was then compared to alternative versions, where regressions techniques were used in order to calculate the model coefficients. The linear regression method revealed to be unusable for the intended purpose, having low performances in all cases. On the other hand, the application of regularization surely brought much better results. In particular, the models obtained using ridge regression were particularly stable and able to have similar performances with the original algorithm.

In the end, the empirical model reduction confirmed to be an approach with great potential. Some of the models were able to simulate the original data with precision even for long periods. However, all the results revealed major problems with the time series regarding temperature anomalies which, in almost all the cases, were poorly represented. Based on some information, like the fact that the pdf of the data is similar to a Gaussian function centered in zero and the general behavior of the data, it is conceivable that the algorithm recognized the signal as a noise and treated it that way. In this situation it is important to reiterate that the temperature data weren't originally intended to be used for model construction.

8.1 Possible future improvements

There are of course many ways in which the project might be further developed, many of them discarded for lack of time or for keeping the analysis and the report more simple. For example, it might be interesting to consider a larger range of possible values for the hyperparameters, such as analyzing models developed up to 20, 30, or even 50 levels. Another possibility is to construct a model not based

on the global evolution of the data but on the hemisphere ones, both at the same time. There are many phenomena present in the two hemispheres which interact with each other and surely some of them might reflect on the models. The last example might be adding other climate data, like the mean precipitation or mean pressure evolution, and see if the algorithm is able to catch relationships with the other components.

As for the algorithm, it might be interesting to build other versions using different methods to calculate the model coefficients, for example, partial least squares regression or principal component regression. Another way might be using neural differential equations or universal differential equations to describe one of the levels, perhaps the main or the last one.

One potential topic is surely time series forecasting. There are already many state-of-the-art models present, like ARIMA, and it might be interesting to compare their performances to the one of the EMR approach.

Finally, using an EMR model as a substitute for part of a general circulation model might help to understand how well the first one is able to work with respect to the second one.

Appendix A

Empirical Model Reduction applied to a Lorenz System

The Lorenz system is a system composed of three ordinary differential equations first studied by Edward Lorenz in 1963 [33]. The three equations (known as Lorenz equations) define a simplified mathematical model for atmospheric convection. They describe in particular the rate of change of three variables. The three equations are:

$$\frac{dx_1}{dt} = \sigma(x_2 - x_1)$$

$$\frac{dx_2}{dt} = x_1(\rho - x_3) - x_2$$

$$\frac{dx_3}{dt} = x_1 x_2 - \beta x_3$$

where x_1 denotes the rate of convection, x_2 the horizontal temperature variations, and x_3 the vertical temperature variations. The constants σ, ρ and β are parameters defined as Prandtl number, Rayleigh number and a geometric factor. Using the set of values $\sigma = 3$, $\rho = 28$ and $\beta = 8/3$, the Lorenz system shows chaotic behaviors, making the system perfect to be used as example of implementing the empirical model reduction as seen in recent publications [12, 8].

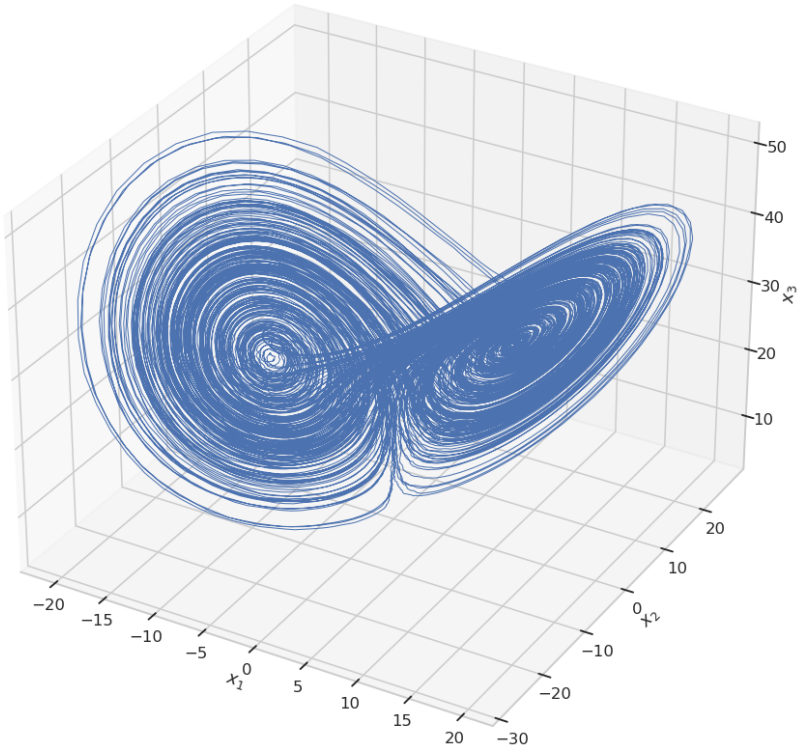


Figure A.1: Trajectories of the Lorenz system in a three-dimensional phase space with $\sigma = 3$, $\rho = 28$ and $\beta = 8/3$.

A.1 Empirical model reduction results

Since the Lorenz equations are quadratics, it will be interesting to analyze the system trying to build both linear and quadratic EMR models. The experiments were conducted on a Lorenz time series of length 300000, with $dt = 0.01$ and sub-sampling with $K = 3$, having at the end 100000 data for each component. Two cases have been studied, one with $eq\_deg = 1$ and one with $eq\_deg = 2$. It has been initially studied, for both cases, the coefficients R^2 on levels up to the 10th one, in order to find the smaller best one.

As shown in Figure A.2, the quadratic model converges to a stable form already after 4 levels, with the best option around 6. This result is easy to imagine, due to the quadratic nature of the Lorenz equations. On the other hand, even with some difficulties, the linear models seem to converge, although with some more levels. To better analyze these results, the best linear case (with 8 levels) and the best quadratic case (with 6 levels) are compared through their autocorrelation and probability distribution functions.

While having some similar behaviors on the first two components, the ACF (Figure A.3) of the EMR model shows that the linear model has more difficulties to maintain the behavior of x_3 than the quadratic one, converging faster to 0.

Examining the probability distribution functions of the components (Figure A.4) it is possible to observe that, even with some imperfections, the quadratic model is able to better cover the value distribution of the original data.

In conclusion, the Lorenz model confirms to be better represented by a quadratic EMR model. However, it is not able to maintain the same behavior of the original data for long, de-correlating after a certain period of time.

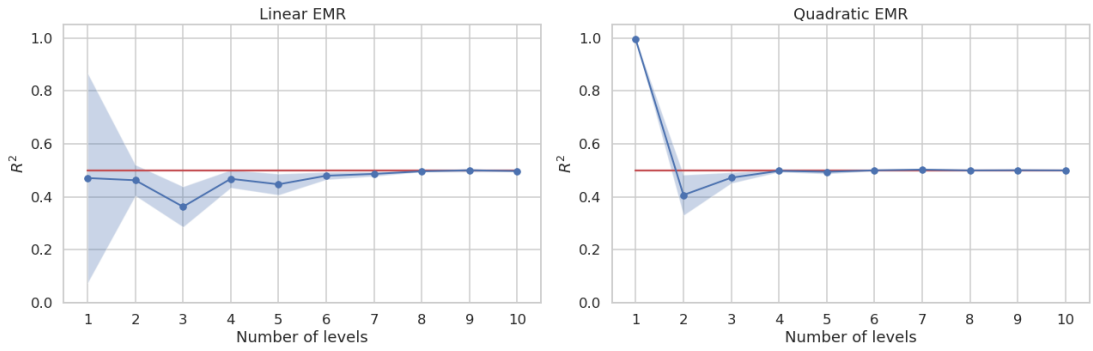


Figure A.2: R^2 coefficients of the linear (left) and quadratic (right) EMR model construction.

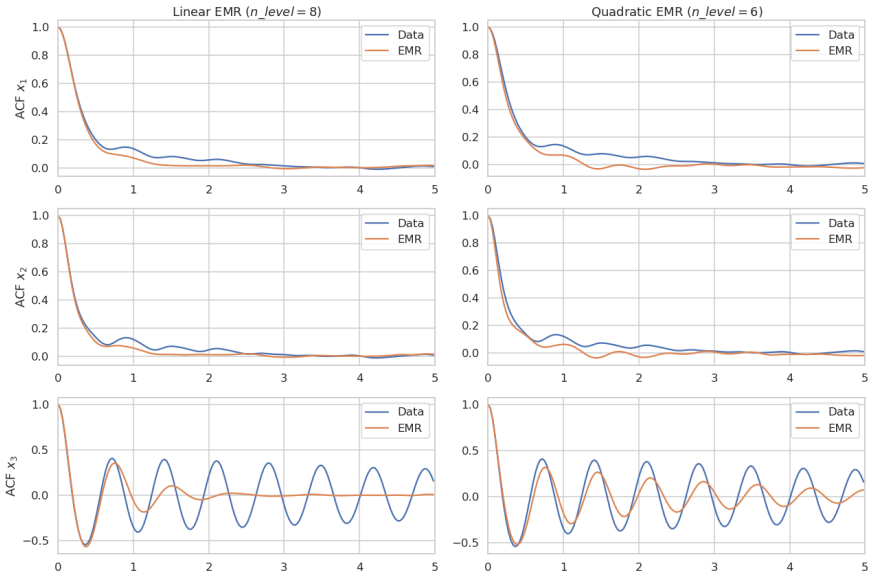


Figure A.3: Comparison of autocorrelation functions between original data and linear (left) and quadratic (right) EMR simulations.

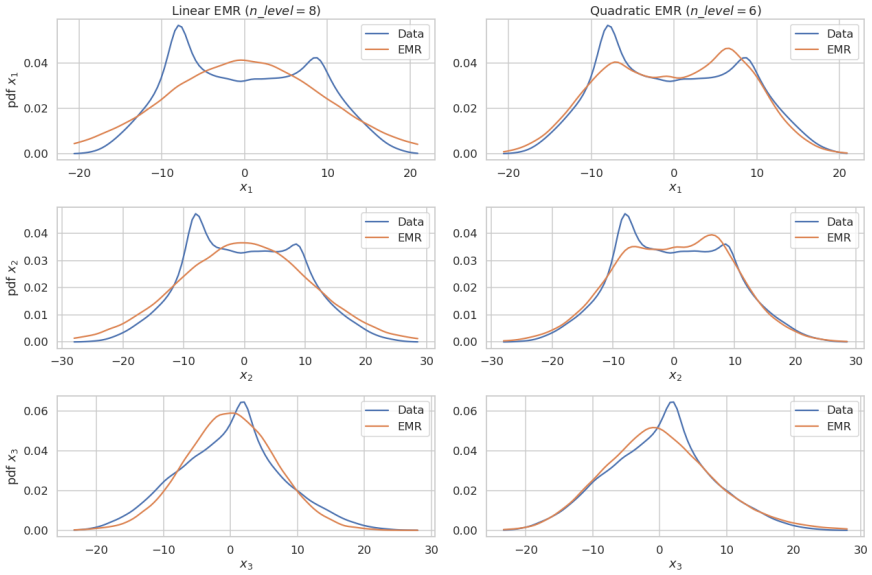


Figure A.4: Comparison of probability density functions between original data and linear (left) and quadratic (right) EMR simulations.

Appendix B

Temperature Anomalies

Temperature anomalies are deviations from the normal temperature for a given location and time of year. In the GISS analysis dataset, the "normal" value always means the average over the 30-year period 1951-1980 for the place and time of year of the measure. Anomalies means are calculated from station anomalies and not from the current absolute mean and the "normal period" mean for that region, but from station temperature anomalies. It must be noted that seasonalities and trends do not depend on the referenced period choice. If the absolute temperature in a specific measure station is higher than a month or a year ago, so are its anomalies regardless of the period chosen.

In general, computing absolute temperature means bringing significant difficulties and large uncertainties [34]. Absolute temperature has significant fluctuations over short distances, while monthly or annual temperature anomalies are able to better represent also larger regions. In general, it has been demonstrated that temperature anomalies are strongly correlated to distances of the order of 1000 km.
Another reason to use temperature anomalies instead of absolute temperature is related to the use of spatial means. Just averaging the available temperatures would give results highly dependent on the particular locations. On the GISTEMP dataset website [26] is present an example that better represents the case.

"Assume, e.g., that a station at the bottom of a mountain sent in reports continuously starting in 1880 and assume that a station was built near the top of that mountain and started reporting in 1900. Since those new temperatures are much lower than the temperatures from the station in the valley, averaging the two temperature series would create a substantial temperature drop starting in 1900."

Appendix C

Data Analysis Images

C.1 Carbon dioxide

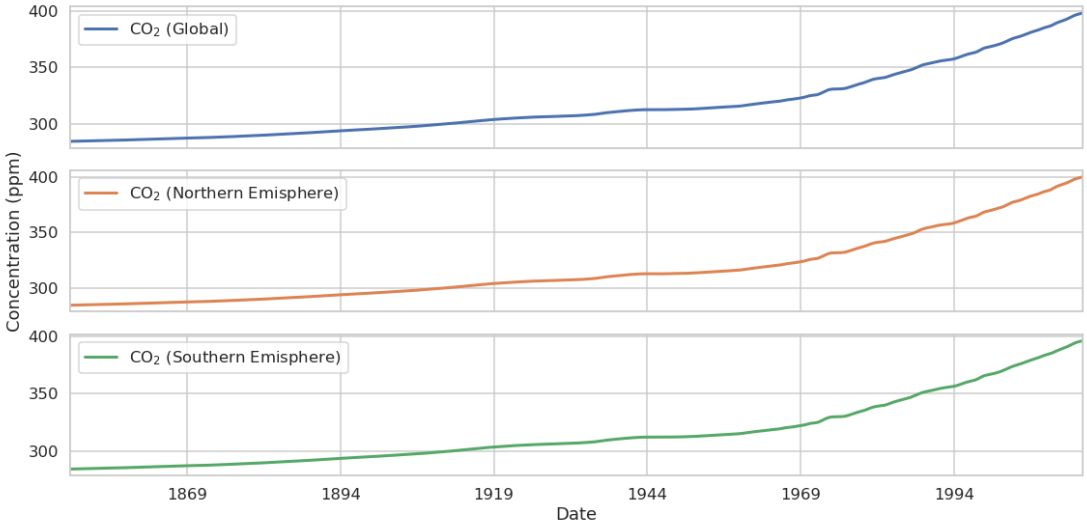


Figure C.1: Trend components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CO_2.

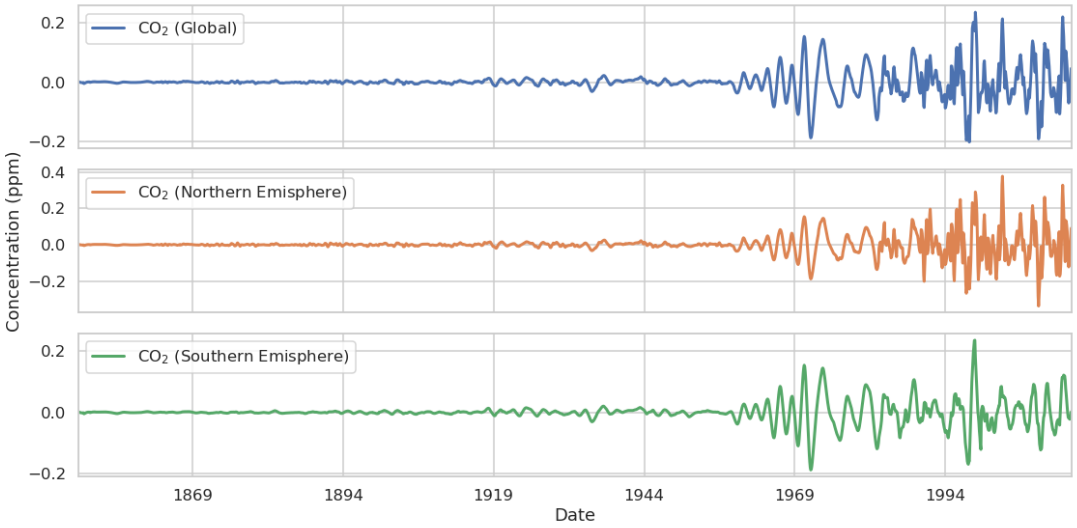


Figure C.2: Noise components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CO_2.

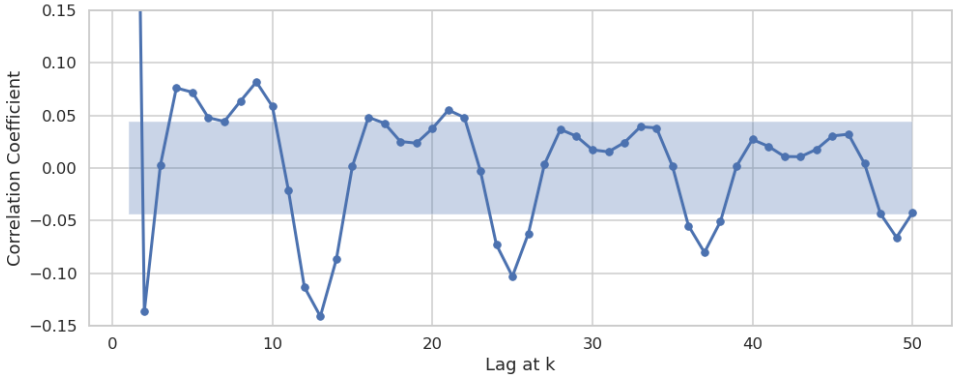


Figure C.3: Partial autocorrelation of CO_2 global concentrations.

C.2 Methane

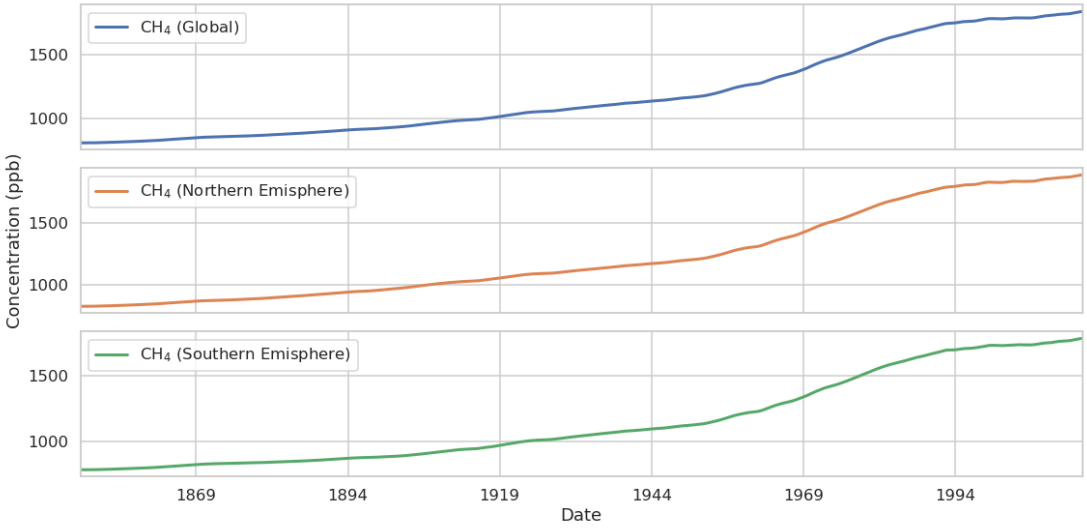


Figure C.4: Trend components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CH_4.

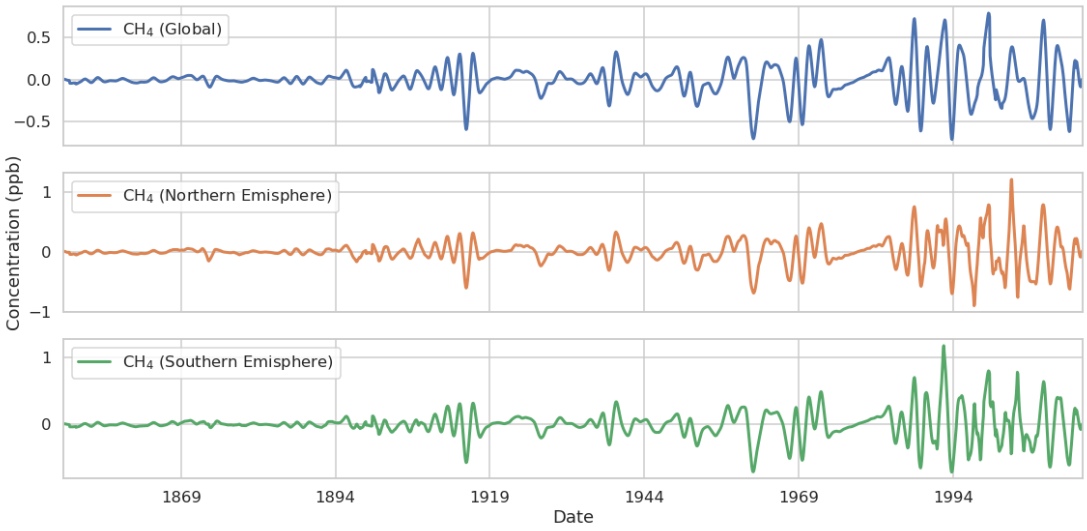


Figure C.5: Noise components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of CH_4.

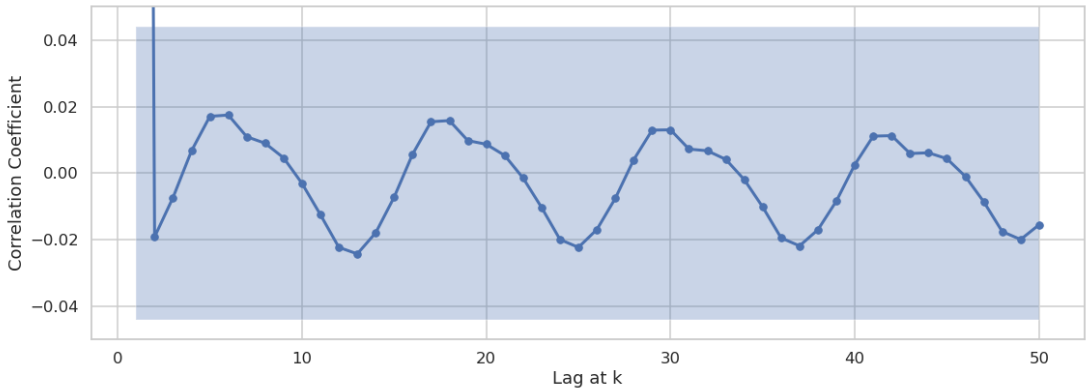


Figure C.6: Partial autocorrelation of CH_4 global concentrations.

C.3 Nitrous oxide

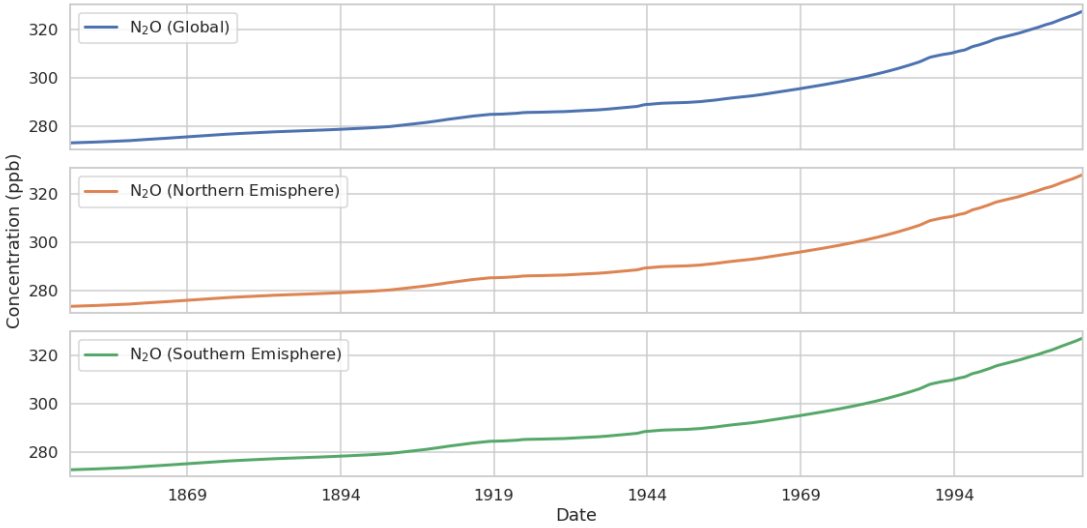


Figure C.7: Trend components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of N_2O.

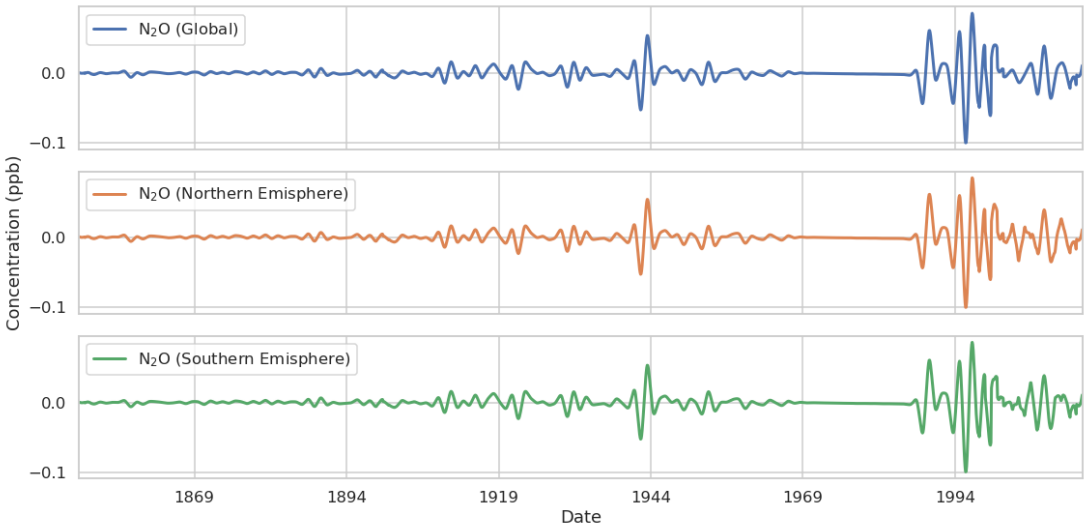


Figure C.8: Noise components of (from top to bottom) global, northern hemisphere and southern hemisphere concentrations of N_2O.

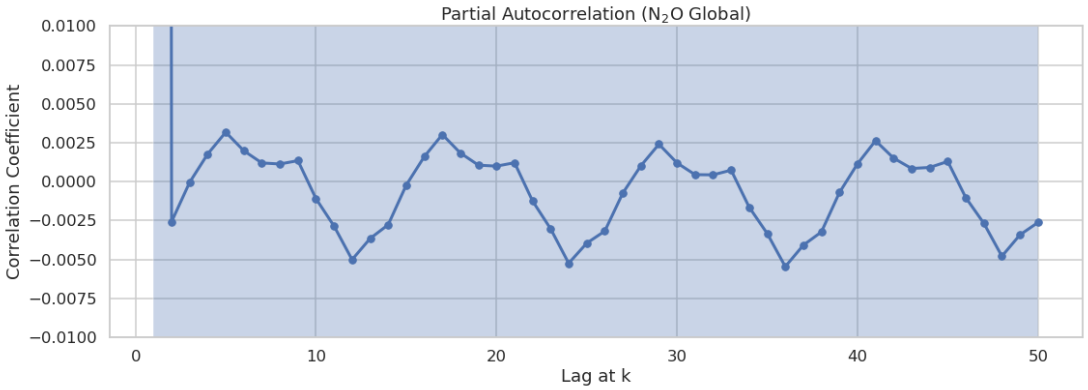


Figure C.9: Partial autocorrelation of N_2O global concentrations.

C.4 Temperature anomalies

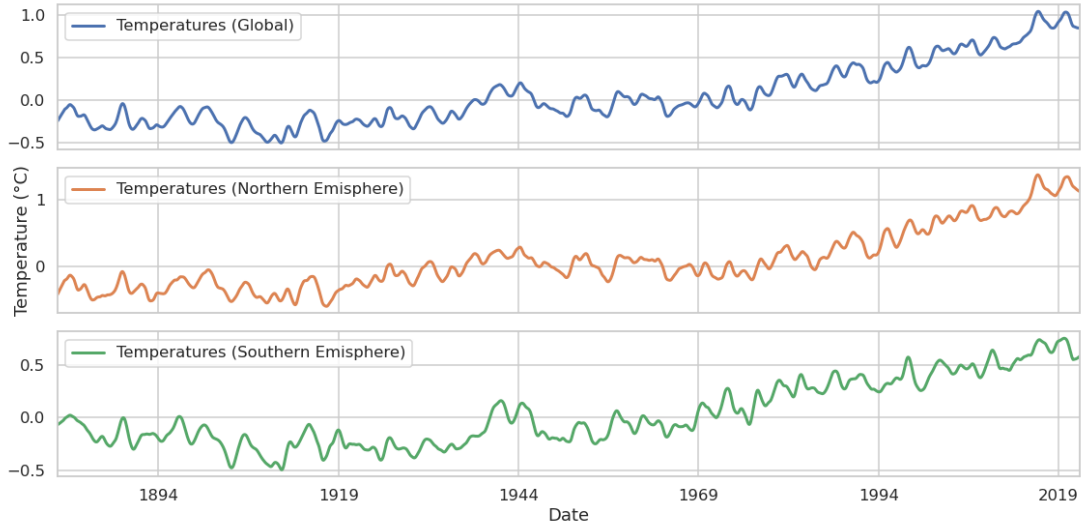


Figure C.10: Trend components of (from top to bottom) global, northern hemisphere and southern hemisphere temperature anomalies.

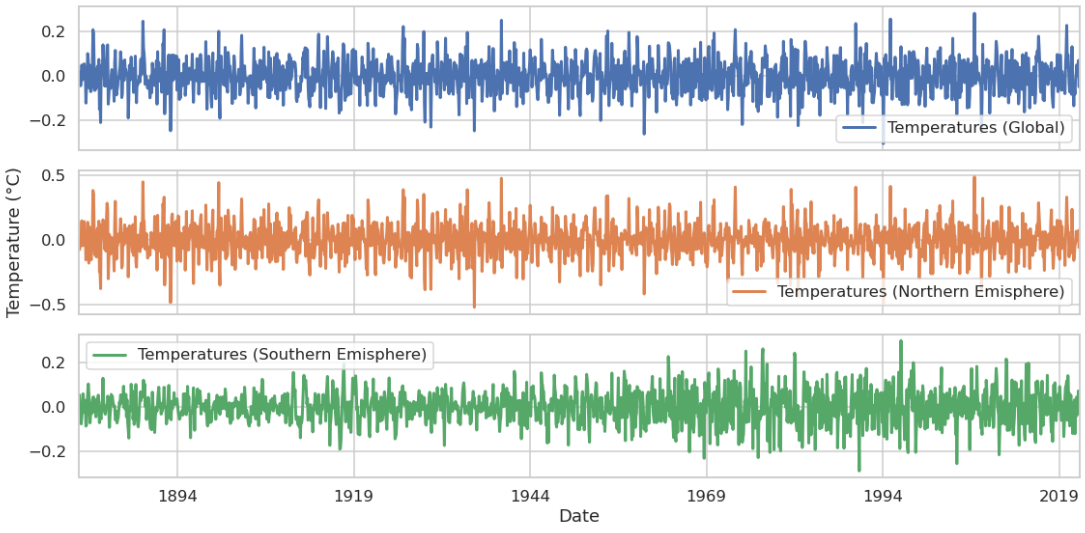


Figure C.11: Noise components of (from top to bottom) global, northern hemisphere and southern hemisphere temperature anomalies.

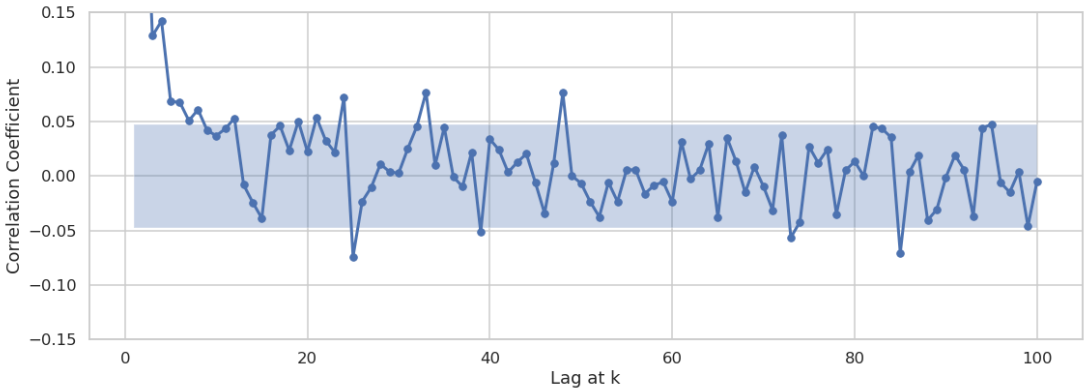


Figure C.12: Partial autocorrelation of global temperature anomalies.

Bibliography

[1] Michael Ghil and Valerio Lucarini. «The physics of climate variability and climate change». In: *Reviews of Modern Physics* 92.3 (2020), p. 035002. DOI: 10.1103/RevModPhys.92.035002 (cit. on pp. 3, 5, 6, 8, 10–12, 14, 16, 17).

[2] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. «Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization». In: *Geoscientific Model Development* 9.5 (2016), pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016 (cit. on p. 3).

[3] Valerio Lucarini, Francesco Ragone, and Frank Lunkeit. «Predicting climate change using response theory: Global averages and spatial patterns». In: *Journal of Statistical Physics* 166.3 (2017), pp. 1036–1064 (cit. on pp. 3, 16).

[4] Matthew Chantry, Hannah Christensen, Peter Dueben, and Tim Palmer. «Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI». In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194 (2021), p. 20200083. DOI: 10.1098/rsta.2020.0083 (cit. on p. 3).

[5] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. «Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere». In: *Journal of Advances in Modeling Earth Systems* 12.9 (2020), e2020MS002109. DOI: https://doi.org/10.1029/2020MS002109 (cit. on p. 3).

[6] Stephan Rasp, Michael S. Pritchard, and Pierre Gentine. «Deep learning to represent subgrid processes in climate models». In: *Proceedings of the National Academy of Sciences* 115.39 (2018), pp. 9684–9689. DOI: 10.1073/pnas.1810286115 (cit. on p. 3).

[7] Paul A. O'Gorman and John G. Dwyer. «Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events». In: *Journal of Advances in Modeling Earth Systems* 10.10 (2018), pp. 2548–2563. DOI: https://doi.org/10.1029/2018MS001351 (cit. on p. 3).

[8] S. Kravtsov, D. Kondrashov, and M. Ghil. «Multilevel Regression Modeling of Nonlinear Processes: Derivation and Applications to Climatic Variability». In: *Journal of Climate* 18.21 (2005), pp. 4404–4424. DOI: 10.1175/JCLI3544.1 (cit. on pp. 3, 19, 117).

[9] D. Kondrashov, S. Kravtsov, and M. Ghil. «Empirical Mode Reduction in a Model of Extratropical Low-Frequency Variability». In: *Journal of the Atmospheric Sciences* 63.7 (2006), pp. 1859–1877. DOI: 10.1175/JAS3719.1 (cit. on p. 3).

[10] Sergey Kravtsov, Dmitri Kondrashov, and Michael Ghil. «Empirical model reduction and the modelling hierarchy in climate dynamics and the geosciences». In: *Stochastic physics and climate modelling* 35 (2009), p. 72 (cit. on pp. 3, 19).

[11] Dmitri Kondrashov, Mickaël D. Chekroun, and Michael Ghil. «Data-driven non-Markovian closure models». In: *Physica D: Nonlinear Phenomena* 297 (2015), pp. 33–55. DOI: 10.1016/j.physd.2014.12.005 (cit. on pp. 3, 19, 20).

[12] Manuel Santos Gutiérrez, Valerio Lucarini, Mickaël D. Chekroun, and Michael Ghil. «Reduced-order models for coupled dynamical systems: Data-driven methods and the Koopman operator». In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31.5 (2021), p. 053116. DOI: 10.1063/5.0039496 (cit. on pp. 3, 117).

[13] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. «Neural ordinary differential equations». In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 3).

[14] Abhinav Gupta and Pierre F. J. Lermusiaux. «Neural closure models for dynamical systems». In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477.2252 (2021), p. 20201004. DOI: 10.1098/rspa.2020.1004 (cit. on p. 3).

[15] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. «Universal Differential Equations for Scientific Machine Learning». In: *arXiv* (2020). DOI: 10.48550/ARXIV.2001.04385 (cit. on p. 3).

[16] Michael Ghil. «A Century of Nonlinearity in the Geosciences». In: *Earth and Space Science* 6.7 (2019), pp. 1007–1042. DOI: https://doi.org/10.1029/2019EA000599 (cit. on p. 5).

[17] Edward N. Lorenz. «Available Potential Energy and the Maintenance of the General Circulation». In: *Tellus* 7.2 (1955), pp. 157–167. DOI: 10.3402/tellusa.v7i2.8796 (cit. on p. 6).

[18] Henry Stommel. «Varieties of Oceanographic Experience». In: *Science* 139.3555 (1963), pp. 572–576. DOI: 10.1126/science.139.3555.572 (cit. on p. 9).

[19] Edward N. Lorenz. «Nondeterministic Theories of Climatic Change». In: *Quaternary Research* 6.4 (1976), pp. 495–506. DOI: 10.1016/0033-5894(76)90022-3 (cit. on p. 13).

[20] David Ruelle. «A review of linear response theory for general differentiable dynamical systems». In: *Nonlinearity* 22.4 (2009), pp. 855–870. DOI: 10.1088/0951-7715/22/4/009 (cit. on p. 15).

[21] Salvatore Caruso, Claudio Giberti, and Lamberto Rondoni. «Dissipation Function: Nonequilibrium Physics and Dynamical Systems». In: *Entropy* 22.8 (2020). ISSN: 1099-4300. DOI: 10.3390/e22080835 (cit. on p. 16).

[22] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013. ISBN: 978-1-421-40794-4 (cit. on pp. 24, 25).

[23] Skipper Seabold and Josef Perktold. «Statsmodels: Econometric and Statistical Modeling with Python». In: *9th Python in Science Conference* (Jan. 2010) (cit. on p. 29).

[24] M. Meinshausen et al. «Historical greenhouse gas concentrations for climate modelling (CMIP6)». In: *Geoscientific Model Development* 10.5 (2017), pp. 2057–2116. DOI: 10.5194/gmd-10-2057-2017 (cit. on p. 31).

[25] N. Lenssen, G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. «Improvements in the GISTEMP uncertainty model». In: *J. Geophys. Res. Atmos.* 124.12 (2019), pp. 6307–6326. DOI: 10.1029/2018JD029522 (cit. on p. 32).

[26] NASA Goddard Institute for Space Studies, GISTEMP Team. *GISS Surface Temperature Analysis (GISTEMP), version 4*. 2022. URL: https://data.giss.nasa.gov/gistemp/ (cit. on pp. 32, 121).

[27] Robert Shumway and David Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Jan. 2017. ISBN: 978-3-319-52451-1. DOI: 10.1007/978-3-319-52452-8 (cit. on p. 33).

[28] James W. Cooley and John W. Tukey. «An algorithm for the machine calculation of complex Fourier series». In: *Mathematics of Computation* 19 (1965), pp. 297–301 (cit. on p. 34).

[29] Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. «STL: A Seasonal-Trend Decomposition Procedure Based on LOESS». In: *Journal of Official Statistics* 6 (1990), pp. 3–73 (cit. on p. 34).

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014. ISBN: 9781461471370. DOI: 10.1007/978-1-4614-7138-7 (cit. on p. 87).

[31] Arthur E. Hoerl and Robert W. Kennard. «Ridge Regression: Biased Estimation for Nonorthogonal Problems». In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.1080/00401706.1970.10488634 (cit. on p. 96).

[32] Robert Tibshirani. «Regression shrinkage selection via the LASSO». In: *Journal of the Royal Statistical Society Series B* 73 (2011), pp. 273–282. DOI: 10.2307/41262671 (cit. on p. 96).

[33] Edward N. Lorenz. «Deterministic Nonperiodic Flow». In: *Journal of Atmospheric Sciences* 20.2 (1963), pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2 (cit. on p. 117).

[34] NASA Goddard Institute for Space Studies, GISTEMP Team. *The Elusive Absolute Surface Air Temperature (SAT)*. 2022. URL: https://data.giss.nasa.gov/gistemp/faq/abs_temp.html (cit. on p. 121).