Università degli Studi di Padova

DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"

Master's Degree in Computational Finance

Bridging meteorology and finance: Italian temperatures for weather derivatives

Supervisor:
Prof.ssa Giorgia Callegaro

Master Candidate: Arianna D'Angelo
Student ID: 2130764

Academic Year 2024/2025

Date 25/07/2025

Abstract

This thesis aims to present a statistical modeling of Italian daily mean temperatures with the aim of applying it to temperature-based climate derivatives. To this end, nine cities were selected in order to best represent the main climate zones of Italy. The first phase consists of an in-depth analysis of the mean, maximum and minimum temperatures observed over the last twenty years. Two different modeling approaches were then developed and compared: a continuous-time model based on an Ornstein–Uhlenbeck process modified to obtain a time-dependent mean and a discrete model, based on an ARMA–GARCH combination applied to the seasonally adjusted residuals. Both models were calibrated on daily historical data and used to simulate, through the Monte Carlo method, the future evolution of the most used temperature indices in the financial field: the Heating Degree Days (HDD) and the Cooling Degree Days (CDD). The simulations obtained were compared with the observed data, evaluating their predictive capacity. Finally, the results were compared with those deriving from the classical burn analysis approach, in order to highlight its advantages and limitations from an application perspective.

Acknowledgments

Desidero ringraziare tutte le persone che mi hanno supportato durante questo percorso, con il pensiero, con l'incoraggiamento o con l'aiuto pratico. Un grazie speciale a Pepe, che con la sua costante supervisione (soprattutto mentre dormiva sulla mia sedia) ha fatto in modo che non lavorassi mai da sola.

Contents

List of Figures

List of Tables

Introduction

Weather and climate have always been a key factor influencing human life. Today, this impact is much less evident than it was a few centuries ago thanks to the advanced technologies we have at our disposal nowadays. From the beginning of civilization to the present day, climate has shaped the organization of work and the lifestyle of many people. This close connection has gradually weakened but has not completely disappeared.

With the discovery of agriculture, about 10,000 years ago, human beings began to lead a sedentary life, based mainly on agricultural activities. The impact of weather conditions on this type of activity was enormous. Just think how a sudden frost at the beginning of spring can damage newly sprouted shoots, compromising the harvest in subsequent seasons. How a lack of rain can drastically reduce crop yields or how an excessively strong wind can break branches and bend cereals, nullifying the efforts of sowing. These are just a few examples of circumstances that, although not catastrophic, can still occur normally and compromise the work of an entire year.

During the Middle Ages, the European economy remained strongly agricultural even if the first urban activities began to develop. Although at first glance the latter seem unrelated to the climate or at least partially independent, it must be considered that commercial activities were mainly carried out outdoors, for example in squares where markets were set up, artisans such as dyers and tanners needed specific conditions to dry their products. In those same years another important component of the economy developed: trade. Merchants, who moved from city to city, needed good weather both to facilitate their travels and to be able to attract customers to the squares where the exchanges took place. Navigation was even more at the mercy of weather conditions and saw a significant development in the Late Middle Ages. Sudden storms or excessively strong winds could slow down travel or damage goods. Very specific conditions were required to complete a sea voyage, such as clear skies, which were necessary for orientation during navigation, winds coming from the right direction and with a strength that was neither too weak nor too strong.

With the Industrial Revolution, the link between weather conditions and work gradually began to weaken. With the spread of factories and the introduction of new technologies, the organization of work began to radically change. Most of the workforce at that time was employed in factories and no longer in the fields. Workers were no longer tied to the hours of natural light, as artificial lighting allowed them to work even outside the hours of daily sunlight. Similarly, weather conditions such as rain or cold no longer directly affected productivity, as work was carried out mainly indoors. Transport also underwent numerous changes. The introduction of the steam engine that led to the first rail transport and the first ships capable of navigating even in the absence of wind or human power made transport increasingly reliable and independent of weather conditions. Goods and people could be transported even over long distances in increasingly shorter times and with greater safety, ensuring continued growth in commercial and industrial activities. This development made the economy less and less vulnerable to climate conditions, reinforcing the separation between work and weather conditions. Today, work seems completely detached from the climate thanks to increasingly advanced technologies that have allowed the creation of controlled environments in which it is possible to carry out activities regardless of external conditions. Imagine how whole plantations inside greenhouses are increasingly spreading throughout the world, where it is possible to control not only the temperature but also the amount of water the plants will receive. However, this work-climate connection has not completely disappeared, temperatures continue to influence energy costs, and transport can be slowed down by rain, fog, and strong winds. Furthermore, tourism is highly dependent on climate conditions: seaside resorts depend on good weather, ski destinations need snowfall or at least temperatures that keep the snow suitable for activities even if artificial. From ancient times to today, the climate has always influenced human life. If once this connection was radical and indissoluble, today it is much less visible, even if not completely eliminated. New technologies have reduced the dependence on weather conditions, but have not completely eliminated it. Man has learned to adapt and manage the influence of climate, but still remains tied to the rhythms dictated by nature.

In any case, the unpredictability of weather conditions continues to represent an economic risk factor for a large number of sectors. Modern activities, despite benefiting from cutting-edge technologies, are certainly not immune to climate fluctuations that can affect energy costs, logistics, demand for goods and services and tourism. Innovative financial instruments known as weather derivatives have been developed to address this uncertainty. These types of instruments were born in the USA in the early 1990s as a response to the needs of companies in the energy sector to protect themselves from fluctuations in energy demand in particularly cool summers or particularly mild winters. Since then, these instruments have evolved and spread across numerous sectors, representing a highly appreciated solution for managing the risk due to weather conditions. The spread of these instruments makes it clear that man, despite having reduced direct dependence on the climate, still seeks to control the economic impact of nature. The analysis of weather derivatives allows us to understand how man, despite having reduced direct dependence on the climate, still seeks to control its economic impact through sophisticated instruments. This study aims to explore the functioning of these tools, their use in different economic sectors, and the opportunities they offer for a more efficient management of climate risk.

This thesis aims to address the topic of weather derivatives applied to the Italian context. The first step is an in-depth analysis of temperatures in Italy by carefully choosing nine cities in order to best represent the different Italian climate zones. This preliminary phase is essential to ensure that the modeling assumptions are consistent with the observed reality. Using the results of the analysis, two models will be developed and compared for the simulation of the HDD and CDD indices: a first model based on a modified Ornstein-Uhlenbeck process and a second based on a combination of ARMA-GARCH. Both models are calibrated on historical data and used to generate forecasts and simulations of the future evolution of the indices, with the final goal of applying them for weather derivatives on temperatures.

The thesis is structured as follows:

- *Chapter 1* presents the context of weather derivatives, their origin, typologies and functioning.
- *Chapter 2* focuses on the analysis of temperature data in Italy.
- *Chapter 3* introduces the two forecasting models used to simulate the HDD and CDD indices, discussing their implementation and the results obtained.
- *Chapter 4* applies models to simulate HDD and CDD indices within the weather derivatives pricing framework, focusing on the comparison between theoretical and numerical approaches.

Chapter 1

Financial market for weather

1.1 History of financial market for weather

"Want a weather forecast? Ask Wall Street."
— By ABC News (February 5, 2010)

A first agreement structured so that the terms depended on the occurrence of specific weather conditions appeared in 1996 and is considered in the literature as the first weather derivative deal. Aquila Energy, a company specialized in the distribution of electricity and natural gas based in Kansas City, Missouri, entered into a contract with Consolidated Edison (ConEd), a company that provides a wide range of energy-related products based in New York, USA. The contract was structured as a dual-commodity hedge for Consolidated Edison (ConEd). The transaction involved ConEd purchasing electricity supplied by Aquila for its August needs. The price of this energy was established at the outset and included a refund if specific conditions regarding average temperatures occurred. The contract stipulated that if the month of August was cooler than expected, in which case the demand for energy would be reduced due to the reduced use of coolers, Aquila would refund ConEd a certain percentage. The definition of a 'cooler than expected August' was certainly not left to chance, but rather specific conditions that could be measured objectively were set. The measurement would therefore be based on the Central Park weather station in New York from which the average daily temperatures would be obtained. If the number of days in which the average daily temperature was below a fixed temperature fell within a certain range, then ConEd would be refunded.

The first weather derivatives started to be traded on the OTC (over-the-counter) market the following year. Noting the market's rapid growth, the Chicago Mercantile Exchange (CME) introduced the first exchange-traded weather derivatives and their associated options in 1999.

In Italy this type of financial product is not widely used because the volumes required in the market for a single farmer are too high. A first weather derivatives contract, as mentioned in [9], in the form of a swap was the one stipulated in 2003 between Banco di Sondrio and Fonte Tavinia, the mineral water company in Salò on Lake Garda. The contract, which took as a reference the temperature detected at Verona-Villafranca airport, provided that Fonte Tavina would pay a predetermined amount for each tenth of a degree of excess of the temperature actually detected compared to the one contractually agreed. In contrast, Banco di Sondrio would pay Fonte Tavina if the detected temperature was lower than the predetermined one since a lower temperature would have lowered the mineral-water company's income because too low temperatures can cause water in pipes or tanks to freeze, causing damage to equipment and production disruptions.

1.2 Weather derivatives and insurance products

The growth of the weather derivatives market is a direct consequence of the growing convergence between the financial and insurance markets. In recent years, there has been a continuous increase in the number of catastrophe bonds issued for several reasons. The first is a consequence of inflation that has hit many world economies, increasing the costs of reconstruction and consequently exposing insurance companies to higher payments in the event of disasters. At the same time, climate change has caused the frequency and severity of extreme weather events, such as hurricanes, forest fires and storms to increase, making losses for insurance companies more significant. As a result, insurance companies have increasingly turned to catastrophe bonds as tools to transfer part of the risk to the financial markets, while simultaneously managing to expand their coverage capacity. Weather derivatives are a natural evolution of these financial instruments as they offer an additional layer of protection against climate risks, allowing companies to transfer exposure to adverse weather conditions to the financial markets rather than relying solely on traditional insurance solutions. Weather derivatives are instruments designed to provide a different choice than an insurance contract to mitigate the risk due to adverse weather conditions. As previously mentioned, weather risk can have a huge impact on many key economic sectors such as agriculture, tourism, transport and energy. Some companies have started to prefer weather derivatives over a classic insurance policy due to the obvious limitations of the latter. An insurance policy provides compensation for a sum of money only if the damage suffered is direct and demonstrable. This very often requires complex and lengthy bureaucratic processes before the money is actually paid out. The situation is different for weather derivatives as payment occurs automatically if a certain measurable meteorological variable, such as millimeters of rainfall, temperature, wind speed, exceeds or does not reach a certain pre-established threshold. In this case, compensation is received by a company without having to prove that it has suffered any direct economic damage so that the liquidation process is quicker and more efficient. Another aspect that makes weather derivatives more advantageous is that, unlike insurance, where contracts are standard and rigidly regulated, weather derivatives, especially those traded on the OTC market, can be structured to suit the two parties, taking into account the different needs that each company may have. From what has been said so far, more and more companies are adopting weather derivatives as an alternative to traditional insurance, as they present a more flexible and effective solution against uncertainty and, consequently, climate risk.

1.2.1 Weather derivative contract

The structure of a weather derivative contract is based on a series of predefined and objective parameters that are monitored over a specific period of time. Each contract is customized based on the specific needs of the company and the type of weather risk that is intended to be covered. A weather derivative contract is made up of the following features:

- Reference location;
- An official weather station from which the measurements are considered;
- Contract period: usually between 1 and 6 months;
- Contract type: future, forward, call, put, swap;
- A weather variable: temperature, average wind's speed, millimeters of rain;
- Threshold: defines the limit value of the meteorological parameter which, if exceeded or not reached, activates the payment provided for in the contract;
- The strike price for the options and future or forward price in the other case;
- Tick that is the amount of money paid out for each unit of the index based on the weather variable;
- Premium in the case of options.

Example 1.2.1. *We consider the contract characteristics reported in Table 1.1. A ski resort in the Dolomites and an airline that flies to and from Bolzano-Dolomiti airport decide to enter into a swap contract on weather derivatives to protect themselves from temperature variations during the winter period, in particular between November and January. The two parties have opposing interests regarding temperatures: the airline fears excessively low temperatures because of the increase in costs due to de-incig procedures and the greater risk of aircraft delays. The ski resort, on the other hand, fears excessively high temperatures that would compromise the quality and quantity of available snow. The contract takes as a reference the average daily temperature measured at the airport weather*

Category	**Value**
Location	Bolzano
Weather station	WMO 16020
Contract type	Swap
Weather variable	Temperature ($°C$)
Contract period	1/11/2023-31/01/2024
Threshold for payment	$4°C$
Strike price (Ski resort)	€100 per degree above $5°C$
Strike price (Airline)	€120 per degree below $4°C$
Measurement frequency	Daily

Table 1.1: Swap weather derivative contract.

station recognized by the World Meteorological Organization, with a threshold of $4°C$ *that represents the approximated to unity average temperature for Bolzano in the period between January 2005 and December 2022 considering only the months of November, December and January. If the temperature exceeds this threshold, that day was milder than expected and the ski resort must resort to artificial snow, increasing costs. In this case, the airline will pay the ski resort a fee of €100 for each degree above* $4°C$*, helping to cover the higher costs of that day. On the contrary, if the temperature drops below that threshold, then the airline will have to face more costs for de-incing operations, therefore the ski resort by paying €120 for each degree below the threshold contributes to these costs, reducing the financial impact on the airline.*

For example, if during a day the average temperature was $7°C$*, the airline would have to pay €300 to the ski resort (€100 for each degree above* $4°C$*). Conversely, if the average temperature was* $2°C$*, the ski resort would pay the airline €240 (€120 euros for each degree below the threshold).*

Using the temperature data from Bolzano for the period 1/11/2023-31/01/2024 we can summarize the results for this contract for which we assume no cost to enter for the two counterparties with the Table 1.2.

Description	**Value**
Days with average temperature $<$ 4°C	44
Days with average temperature $\geq$ 4°C	48
Sum of temperature excess above 4°C	107
Sum of temperature shortfall below 4°C	98
Total money received by airline	11,760
Total money received by ski resort	10,700

Table 1.2: Results of the exchange based on temperature.
Data source: ilmeteo.it

This contract allows both parties to reduce uncertainty related to winter weather con-

ditions, while protecting their respective economic interests.

1.3 Different types of weather derivatives

In this section we present a classification of the different categories of weather derivatives present on the financial market. The categories are taken from [2].

1.3.1 Temperature derivatives

Temperature derivatives, the largest weather derivatives traded on the market, were designed to try to protect businesses and investors from temperature fluctuations that could affect expected profits. They are widely used in sectors where weather plays a key role, as it significantly affects the supply and demand for goods and services.

As mentioned above when we talked about the first temperature-based contract, the use of temperature derivatives can be very useful in the energy sector by helping companies manage fluctuations in demand and hedge against the risk of an unexpected drop in demand. A natural gas supplier could experience serious losses if a winter is much milder than expected because demand for gas, used primarily for heating, would see a drastic drop compared to expectations. In this case, the gas supplier could hedge itself by entering into a contract that allows it to receive financial compensation if the winter temperatures exceed a certain threshold to mitigate losses.

Another sector that is highly dependent on weather conditions, and especially temperatures, is agriculture. If a summer is particularly cold, crop growth could slow down, decreasing crop yields in subsequent seasons. In this case, a contract that provides compensation in the event of lower temperatures than expected could compensate for the loss of earnings from a poor harvest.

Undoubtedly, one of the sectors that is vulnerable to temperatures is tourism. A ski resort facing a warmer winter than expected, perhaps considering previous seasons, could risk a huge loss if the snow melts. Similarly, beach resorts can see their revenues decrease in the event of a cool summer.

Temperature derivatives are at the center of the discussion of this thesis; therefore in the paragraph 1.4 there will be a detailed discussion of such derivatives.

1.3.2 Derivatives on wind speeds

Wind speed derivatives are instruments used to manage the risk associated with changes in wind speed, which can affect various economic sectors.

With the increasing use of renewable sources for energy production, including wind energy, wind speed is a key factor in determining the productivity of wind turbines, as well as in planning and managing the risks associated with energy generation. If the wind is too weak, turbines produce a reduced amount of energy, consequently reducing revenues. Conversely, if the wind speed is too high, wind farm operators must close down the mills in order to avoid damaging the rotors. A company that operates wind farms can enter into a contract that protects it from a period of isolated weak or excessively strong winds in order to provide financial compensation to cover the losses due to the lack of energy production.

Another sector that could benefit from wind speed derivatives is the aviation sector. Airlines and airport operators can hedge against the risk of cancellations, delays and increased fuel consumption due to the need for more power during take-off due to strong winds near runways.

Nordix wind speed index

In 2007 the US Futures Exchange (USFE) decided to open a market for wind speed derivatives. Although it was closed shortly after, at the end of the 2008, its market specification provided a natural framework for a wind speed derivatives; therefore, the specification is discussed below. The underlying of USFE futures contracts was the so-called Nordix wind speed index (from now on we will refer to it as the Nordix index for simplicity). The measurement of this index was the daily average wind speed collected in two wind farms, one in New York state and the other in Texas. The index is based on the deviations of the daily wind speed from a 20-year mean usually aggregated over a measurement period like a month or a season. Additionally, a benchmark value of 100 was introduced. Let $W(t)$ be the average daily wind speed measured on day t, $w_{20}(t)$ be the mean of the last 20 years', then the Nordix index over a period from $\tau_1 \in \mathbb{N}$ to $\tau_2 \in \mathbb{N}$ with $\tau_1 < \tau_2$ is defined as

$$N(\mathcal{T}) = 100 + \sum_{t=\tau_1}^{\tau_2} (W(t) - w_{20}(t))$$

with $\mathcal{T}$ the set of the days considered $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$. Therefore, if $N(\tau_1, \tau_2) > 100$, wind farms produced more than expected based on the 20-year mean assumption. In the other case if $N(\tau_1, \tau_2) < 100$ then the production was less than expected. This uncertainty can be hedged by trading futures on the Nordix index, essentially swapping a floating index for a fixed one.

1.3.3 Precipitation derivatives

These types of derivatives are based on data such as the level of precipitation, rain or snow as established, recorded in a given area and in a given period. Their use is particularly useful for industries that depend on weather conditions, such as agriculture, energy and tourism.

For example, as previously mentioned, the agricultural sector is highly vulnerable to weather conditions. In the specific case of precipitation, excessive or insufficient rain can compromise the harvest. Using precipitation derivatives, a farm can protect itself from the risk of drought or on the contrary from excessive rainfall by entering into a contract that provides compensation in the event that precipitation falls or exceeds a pre-established threshold in order to try to compensate for losses.

Another sector that can use this type of financial derivatives is the energy sector, especially companies that manage hydroelectric plants for which it is important to have at least a certain level of rainfall so that there is a sufficient level of water for energy production. These companies can enter into a contract that in the case of particularly dry seasons can financially cover the losses due to the lack of energy production.

In tourism and transportation, snow is a determining factor. Ski resorts can use snow derivatives to protect themselves against winters with lower than average snowfall, which could reduce the number of visitors and therefore profits. Similarly, transportation companies can hedge against the risk of excessive snowfall that could cause delays or interruptions in service.

For several years, the CME has offered snowfall futures for some U.S. cities as well as options on precipitation. The basis of showfall futures is the CME Snowfall Index, which measures the amount of snow that falls over a predefined measurement period. Showfall futures are traded from November to April in monthly aggregates or in blocks of several months. The CME Snowfall Index over a measurement time period $[\tau_1, \tau_2]$ is defined as aggregated daily snowfall in inches over the period. Let $S(t)$ be the total inches of snow on day t with $t \in \mathcal{T}$ defined as $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$, the Snowfall Index is defined as

$$SF(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} S(t).$$

For rainfall the discussion is similar with the CME Snowfall Index which measures in inches the amount of rain that fell in a specific place in a specific time interval. The Rainfall Index is defined as

$$RF(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} R(t)$$

with $R(t)$ the amount of rain in inches of rain for the day t.

1.3.4 Frost derivatives

There are many other categories of weather derivatives on the market, although less widespread and known. An example could be the derivative offered by CME on the Frost Index. The Frost Index, usually traded as futures, measures the days in which there is frost on weekdays from November to March at the Amsterdam Schipol airport. Ice in an airport is very dangerous. On the aircraft, frozen components can cause a lot of damage: ice can lead to an alteration of the aerodynamics by reducing lift and increasing aerodynamic resistance compromising flight capabilities, accumulations of ice can prevent the correct functioning of mobile surfaces such as flaps and rudders, ice in the air intakes reduces the efficiency of the engines and finally it can obstruct vital flight sensors such as the pitot tubes which, by reading incorrect airspeed values due to ice, can lead to a loss of control. The same ice can also be dangerous on runways: reduced grip as the runway is more slippery, consequently increasing take-off and landing distances as well as making it difficult to maneuver the aircraft. Precisely because of these imminent risks as soon as temperatures approach zero, airlines and airports must follow a series of practices such as de-icing which, by raining antifreeze on the aircraft, allows the surfaces to be de-iced. This practice must be carried out shortly before take-off and if the maximum waiting time between de-icing and take-off is exceeded, which is very easy in very busy airports such as Schipol, the aircraft must re-queue for de-icing and take-off is delayed. That said, some airlines may want to protect themselves from the costs of delays by entering into a contract that allows them to receive compensation if the days with icy runways exceed a certain limit.The CME Frost Index calculates the number of days in a given month or the entire season from November to March in which frost takes place, according to specific temperature qualifications.

1.4 Temperature derivatives

The reference market for trading in financial derivatives is the CME, therefor in this thesis the approach to this particular financial derivatives used is the one proposed by the CME. The most easy financial derivative we can define on a particular index is a future. The futures traded are settled against three main indices: cooling degree-day, heating degree-day and cumulative average temperature. Before proceeding to give the definition, taken from [2] and [18], of this three indices and of other lesser known indices is important to give this definition

Definition 1.4.1. *The average temperature for a given day t is defined as the mean of the recorded maximum and minimum temperature in that day*

$$T(t) = \frac{T_{min}(t) + T_{max}(t)}{2}.$$

The following two indices are both measured against a benchmark of 65° Fahrenheit. However, since the temperatures considered in this work are in Italy, they will be measured in degrees Celsius. Therefore, the benchmark temperature will be $18°C$.

Definition 1.4.2. *The cooling degree-day (CDD from now on) index is defined as the positive part of the difference between the average temperature for a specific day t and the* *benchmark of* $18°C$

$$CDD(t) = max\,(T(t) - 18, 0) = (T(t) - 18)^+.$$

The index structured in this way can be a useful and simple factor for estimating energy demand. The basic logic is that high temperatures, which imply a positive index, are generally associated with a higher energy demand due to the use of coolers. Typically, futures are not settled based on the index of a single day but rather on the aggregate value of the indices over a specific sequence of consecutive days. The days selected are commonly referred as measurement period that is typically a week, a month or a sequence of months, from two to seven, that is referred as seasonal strip. Generally the summer seasonal strip is defined from April to October. In this case it is convenient to define the index in an alternative way to make the definition more general by including also the possibility to add a strip.

Definition 1.4.3. *The cumulative CDD index can be defined as*

$$CDD(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau 2} (T(t) - 18)^+$$

with $t \in \mathcal{T}$ *and* $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$.

Definition 1.4.4. *The heating degree-day (HDD from now on) index is defined as the positive part of the difference between the benchmark of* $18°C$ *and the average temperature for a specific day* t

$$HDD(t) = max\,(18 - T(t), 0) = (18 - T(t))^+.$$

Similarly, the HDD index uses temperature as a simple and effective indicator to estimate energy demand. When temperatures fall below the benchmark, the need for heating in buildings increases, resulting in higher energy consumption. The HDD index measures how much the daily temperature is below this reference value, thus providing a useful tool for forecasting and managing heating energy demand. Also in the case of the HDD index is possible to propose a more general definition. A common choice for the measurement period is the winter seasonal strip that aggregate the daily temperature from November to March.

Definition 1.4.5. *The cumulative HDD index can be defined as*

$$HDD(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau 2} (18 - T(t))^+$$

with $t \in \mathcal{T}$ *and* $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$.

To obtain a general indication of the energy demand regardless of whether for heating or cooling, it is possible to use the following index

Definition 1.4.6. *The energy degree days (EDD) index from τ_1 to τ_2 is defined as the sum of the CDDs and HDDs over the interval $[\tau_1, \tau_2]$,*

$$\begin{aligned} EDD(\tau_1, \tau_2) &= CDD(\tau_1, \tau_2) + HDD(\tau_1, \tau_2) \\ &= \sum_{t=\tau_1}^{\tau_2} (18 - T(t))^+ + (T(t) - 18)^+ = \sum_{t=\tau_1}^{\tau_2} |18 - T(t)| \end{aligned}$$

with $t \in \mathcal{T}$ and $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$.

The following index defined is the one that takes into account the average daily temperatures regardless of whether they exceed or fall below a certain threshold.

Definition 1.4.7. *The cumulative amount of temperature (CAT) index from τ_1 to τ_2 is defined as the sum of the average daily temperature over the interval $[\tau_1, \tau_2]$,*

$$CAT(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} T(t)$$

with $t \in \mathcal{T}$ and $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$.

Plants require a period of temperatures sufficiently high to allow them to grow. It is possible to empirically define for some species a minimum threshold below which the growth of the plant is compromised. For example evergreen conifers such as of Picea and Abies, by cold-or summergreen trees like Betula, Populus or Larix need a minimum temperature of $5°C$. This need can be represented by the index of accumulated growing-season warmth (GDD) or as usually referred to in the financial sector as Growing degree days:

Definition 1.4.8. *The Growing degree days (GDD) index from τ_1 to τ_2 is defined as the sum of the difference between the daily temperature mean and the minimum temperature for a correct plant growth T_0 over $[\tau_1, \tau_2]$,*

$$GDD(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} (T(t) - T_0)^+$$

with $t \in \mathcal{T}$ and $\mathcal{T} = \{t \in \mathbb{N} \mid \tau_1 \leq t \leq \tau_2\}$.

Example 1.4.1. *Using the average temperature data for nine of the largest Italian cities for the week of May 17-23, 2010 in the Table 1.3, it is possible to calculate the CDD index value for the individual city. The data are collected from* `ilmeteo.it`. *An asterisk (*) is used to indicate numbers below $18°C$. The values are reported in the Table 1.3. To*

Date	Bari	Ancona	Torino	Palermo	Viterbo	Roma	Venezia	Bolzano	Bologna
17-05-2010	15*	16*	15*	17*	15*	14*	18	17*	18
18-05-2010	16*	17*	16*	17*	17*	16*	18	17*	18
19-05-2010	16*	15*	15*	17*	14*	16*	17*	17*	14*
20-05-2010	14*	15*	17*	17*	16*	16*	18	19	17*
21-05-2010	16*	17*	18	18	17*	17*	19	20	19
22-05-2010	18	18	20	18	19	17*	20	20	21
23-05-2010	19	18	21	18	21	18	21	21	22

Table 1.3: Mean temperature for the week from 17/05/2010 to 23/05/2010.

calculate the index value, proceed as follows:

$$CDD(17/05/2010, 23/05/2010) = \sum_{t=17/05/2010}^{23/05/2010} (T(t) - 18)^+.$$

If we take the temperatures of the city of Bologna as a reference then

$$CDD(17/05/2010, 23/05/2010) = 2*(18-18)^+ + (19-18)^+ + (21-18)^+ + (22-18)^+ = 8.$$

The Table 1.4 shows the index values for the selected week and the number of days in which the temperature is above $18°C$

	Bari	Ancona	Torino	Palermo	Viterbo	Roma	Venezia	Bolzano	Bologna
Days with T>=18°C	2	2	3	3	2	1	6	4	5
Total sum of exceeds	1	0	5	0	4	0	6	8	8

Table 1.4: CDD index values for the week from 17/05/2010 to 23/05/2010.

Example 1.4.2. *Using the average temperature data for the same cities as in the previous example but this time for the week of October 10-16, 2010 in the Table 1.5 it is possible to calculate the HDD index value for the individual city. An asterisk (*) is used to indicate numbers below* $18°C$*. To calculate the index value, proceed as follows:*

Date	Bari	Ancona	Torino	Palermo	Viterbo	Roma	Venezia	Bolzano	Bologna
10-10-2010	14*	13*	14*	22	15*	18	13*	14*	12*
11-10-2010	14*	14*	13*	23	12*	16*	13*	13*	12*
12-10-2010	16*	15*	12*	22	16*	19	14*	12*	13*
13-10-2010	17*	15*	12*	21	16*	19	13*	11*	12*
14-10-2010	17*	14*	13*	20	16*	17*	15*	11*	14*
15-10-2010	17*	15*	13*	20	15*	16*	13*	11*	13*
16-10-2010	16*	14*	12*	21	14*	16*	12*	10*	12*

Table 1.5: Mean temperature for the week from 10/10/2010 to 16/10/2010.

$$HDD(10/10/2010, 16/10/2010) = \sum_{t=10/10/2010}^{16/10/2010} (18 - T(t))^+.$$

If we take the temperatures of the city of Rome as a reference then

$$HDD(10/10/2010, 16/10/2010) = 3 * (18 - 16)^+ + (18 - 17)^+ = 7.$$

The Table 1.6 shows the index values for the selected week and the number of days in which the temperature is under $18°C$.

	Bari	**Ancona**	**Torino**	**Palermo**	**Viterbo**	**Roma**	**Venezia**	**Bolzano**	**Bologna**
Days with T<18°C	7	7	7	0	7	4	7	7	7
Total sum of exceeds	15	26	37	0	22	7	33	44	38

Table 1.6: HDD index values for the week from 10/10/2010 to 16/10/2010.

1.5 CME

The CME Group (Chicago Mercantile Exchange Group) is one of the largest derivative exchanges in the world. Formed from the merger of four major exchanges, the Chicago Mercantile Exchange (CME), the Chicago Board of Trade (CBOT), the New York Mercantile Exchange (NYMEX), and the Commodity Exchange Inc. (COMEX), the CME Group provides a wide range of financial instruments such as futures and options on commodities, interest rates, currencies, stock indices, and energy, as well as more innovative products. In addition to being a leader in futures and options, the CME Group revolutionized the market with the introduction of weather derivatives, providing traders with tools to mitigate the risks associated with adverse weather events. These instruments are particularly useful for sectors such as energy, agriculture and tourism, which are highly influenced by weather conditions. The weather derivatives offered by the CME Group are distinguished by the duration, which can be monthly or seasonal, typically covering three to five months. The standard reference seasons are the heating season, from November to March, and the cooling season, from May to September. This contracts are based on daily temperature measurements and are adjusted based on deviations from a reference value, generally $65°F$ ($18°C$). The products offered by the CME Group vary depending on the continent. In North America, the most common contracts are based on Heating Degree Days and Cooling Degree Days. In Europe, in addition to HDDs, the Cumulative Average Temperature index is often used for the summer months. In Asia, and therefore in Tokyo, all weather contracts refer exclusively to the CAT Index, regardless of the season. In addition to temperature-related contracts, the CME Group offers instruments to hedge risks related to other meteorological phenomena. For New York and Boston, futures and options on the Snowfall Index are available, which allow you to manage the financial impact of abnormal snowfall. Hurricane Seasonal futures and options are offered in territorial bands of the USA. In addition, for Amsterdam, there is a Frost Index, designed to protect agricultural companies and sectors particularly exposed to the risk of frost, such as airlines.

The Table 1.7 reports the main cities for which futures and options contracts are offered by the CME Group for a specific index. The data are taken from [6] and [7].

City	**Continent**	**HDD**	**CDD**	**CAT**
Atlanta	America	✓	✓	×
Boston	America	✓	✓	×
Burbank	America	✓	✓	×
Chicago	America	✓	✓	×
Cincinnati	America	✓	✓	×
Houston	America	✓	✓	×
New York	America	✓	✓	×
Dallas	America	✓	✓	×
Las Vegas	America	✓	✓	×
Minneapolis	America	✓	✓	×
Sacramento	America	✓	✓	×
Philadelphia	America	✓	✓	×
Portland	America	✓	✓	×
Essen	Europe	✓	×	✓
London	Europe	✓	×	✓
Amsterdam	Europe	✓	×	✓
Paris	Europe	✓	×	✓
Tokyo	Asia	×	×	✓

Table 1.7: Availability of weather contracts by city.

Chapter 2

Temperature analysis of Italian cities

"The common cliché is that every conversation begins and ends with the weather."
— By CME Group (Weather Derivatives Brouchure)

2.1 Introduction to the analysis of historical temperature data for derivative pricing

The following chapter aims to provide an analysis of the historical temperatures of the last twenty years for some of the main Italian cities. Through the selected cities, Turin, Bologna, Bolzano, Venice, Viterbo, Rome, Ancona, Bari and Palermo, we want to try to provide a clear overview of the main climatic zones into which the Italian peninsula can be divided. The analysis of such historical data will be essential to better understand the temperature pattern to define in the best possible way a model that describes its course. The definition and the estimation of the model will be necessary to subsequently move on to the pricing of temperature derivatives, increasingly widespread financial instruments that are linked to climatic conditions such as extreme temperatures or seasonal averages above or below a certain level. For the pricing of these derivatives, understanding the historical dynamics of temperatures in the areas of interest is essential. Identifying historical trends will be essential to build models that best replicate the temperature patterns. Temperature analysis, combined with statistical modeling, will provide the basis for developing a robust and reliable model for pricing weather derivatives, which can be adapted to the different climatic and morphological characteristics of Italian cities.

2.2 Data sources for the analysis

The data regarding maximum, minimum and average daily temperatures used in the following analysis come from the historical archive of the well-known weather forecasting website ilmeteo.it, which provides detailed meteorological information for various locations in Italy. This data is accessible via the following link: `https://www.ilmeteo.it/portale/archivio-meteo`. For the individual cities selected, the data comes from the local reference meteorological stations that guarantee precise and representative measurements of the climatic conditions. The meteorological stations from which the data are collected are indicated below for each city.

- Turin: the data comes from the Turin-Caselle meteorological station, located at the Turin-Caselle Airport managed by ENAV (Ente Nazionale Assistenza al Volo). The station, owned by the Italian Air Force Meteorological Service, is also recognized by the World Meteorological Organization (WMO 16059);

- Bolzano: the data for the city of Bolzano come from the Bolzano San Giacomo Meteorological Station, located at the Bolzano Dolomiti Airport. This station is operationally managed by ENAV for air traffic needs, but belongs to the Italian Air Force Meteorological Service. Also this station is recognized by the World Meteorological Organization (WMO 16020);

- Venice: the reference weather station for Venice is in Venice-Tessera, located at Venice Marco Polo Airport. Managed by ENAV but it belongs to the Italian Air Force Meteorological Service. The station is recognized by the World Meteorological Organization (WMO 16105);

- Bologna: the reference station for Bologna is Bologna-Borgo Panigale, located at the Bologna Guglielmo Marconi Airport managed by ENAV. The station is owned by the Italian Air Force Meteorological Service. The station is recognized by the World Meteorological Organization (WMO 16140);

- Viterbo: the data comes from the Viterbo Aeroporto Meteorological Station, located at the Viterbo airport at about 300 meters above sea level. This station is managed by the Italian Air Force Meteorological Service and is recognized by the World Meteorological Organization (WMO 16216);

- Rome: the data for Rome are collected by the Ciampino weather station located at the Rome-Ciampino airport. It is managed by ENAV but belongs to the Italian Air Force Meteorological Service. The station is recognized by the World Meteorological Organization (WMO 16239);

- Ancona: the reference weather station, Ancona Falconara, is located in the municipality of Falconara Marittima and belongs to the Italian Air Force Meteorological Service, although managed by ENAV. The station is recognized by the World Meteorological Organization (WMO 16191);

- Bari: the data comes from the Bari-Palese Meteorological Station, located at the Karol Wojtyła Airport. The station is operationally managed by ENAV for air navigation, but belongs to the Italian Air Force Meteorological Service, which is responsible for the collection and distribution of meteorological data. The station is recognized by the World Meteorological Organization (WMO 16270);

- Palermo: the data comes from the Palermo-Punta Raisi Weather Station, located at the Falcone-Borsellino Airport. Although the operational management of the station is entrusted to ENAV, the station belongs to the Italian Air Force Meteorological Service, which is responsible for the collection and dissemination of meteorological data. The station is recognized by the World Meteorological Organization (WMO 16405).

Each weather station provides a data set that accurately reflects local climate conditions, allowing for an in-depth understanding of historical temperatures and their variability in different Italian cities. In this context, it is essential to specify the reference meteorological stations precisely. The choice to use WMO-recognized stations guarantees high standards of quality and continuity in the observed data, essential elements to ensure the transparency and reliability of contracts based on climate parameters.

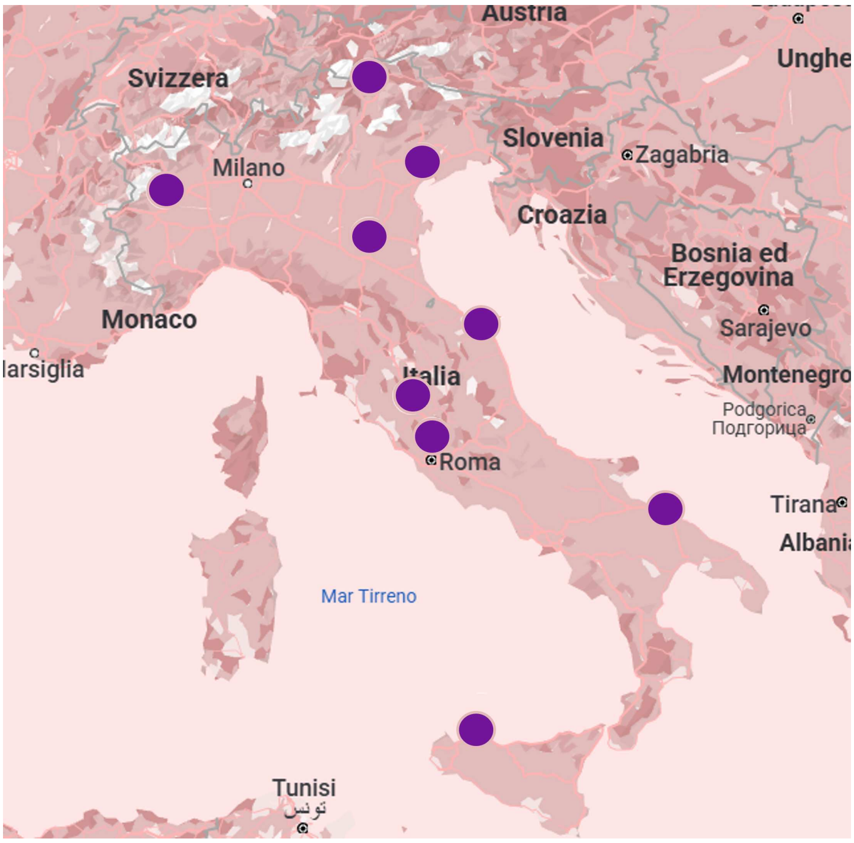


Figure 2.1: Locations of selected Italian cities for temperature data analysis, marked by indigo points.

2.3 Factors that influence the climate

Weather is constantly changing both in time and space. The cause of these variations is the different combination in quantity and intensity of climatic factors such as air temperature, pressure, wind, humidity and precipitation. There are two categories of factors that influence these climatic elements, cosmic and geographical. The Figure 2.2, inspired by the diagram in [16], aims at representing what will be explained in the current paragraph.

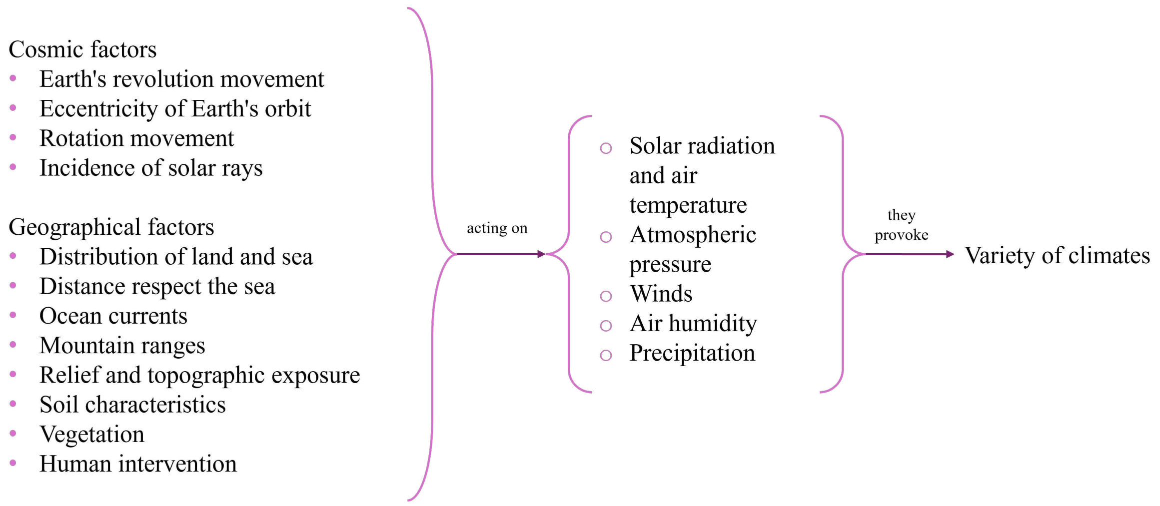


Figure 2.2: Diagram of the factors and elements of the climate.

2.3.1 Cosmic factors

Phenomena that occur in the Earth's atmosphere draw energy from the sun. The solar radiation that strikes a specific point on the Earth's surface depends on the particular moment in which it occurs. Factors such as the Earth's position in its orbit around the sun or its rotation significantly influence climate elements such as air temperature. The cosmic factors that influence the climate are:

- The Earth's revolution movement;
- The eccentricity of the Earth's orbit;
- The rotation movement;
- The incidence of solar rays.

2.3.2 Geographical factors

Cosmic factors follow a very specific pattern that would allow the earth to be divided into climate zones using latitude. According to this potential subdivision, places that are in the same parallel should have the same climate. In reality, this is almost never observed. A classic example of this are the two cities of Naples, Italy, and New York, USA, which are both north parallel 41°, but have significantly different climates. In the literature, climatologists define the set of factors on the Earth's surface that influence climate conditions. These are physical factors that hinder or favor specific climate elements. The main geographical factors that influence climate conditions, such as temperature, the object of the study, will be listed below.

Distribution of land and sea

Looking at a planisphere it is possible to conclude that the lands tend to group together in large masses, the continents, which remain separated from each other by large bodies of water, the oceans. The difference between land and sea distribution has a great influence on climates, an example is the widespread continental climate in much of the Eurasian continent, but little in the southern hemisphere where the percentage of newly deposited land slightly exceeds 16%. The different behavior based on the percentage of land and sea is mainly due to the different behavior with respect to solar radiation, as the seas tend to heat up and cool down more slowly than the land for two reasons: the first is the specific heat of water, which is approximately double that of the land, and the second is that the land is more opaque than the ocean as it is capable of absorbing solar rays only in the most superficial layers.

Distance respect the sea

The areas near the sea, precisely because of the characteristics already mentioned, have an average excursion much lower than those that occur within the continents. The sea, in addition to attenuating the oscillations of diurnal and annual temperatures, causes seasonal extreme temperatures, such as cold in winter and heat in summer, to be delayed with respect to the solstices to which they are linked. In seaside locations, there is another effect due to their proximity to the body of water: winters tend to be longer. An example with the data at our disposal can be seen with the two Italian cities of Venice and Bologna. The first has an average January temperature in 2005-2024 of $3.85°C$ while the second of $3.48°C$. However, in March, the situation is very different, with Venice having an average temperature of $9.21°C$ while Bologna having $9.75°C$.

Ocean currents

Ocean currents have a decisive influence on the regional distribution of climates. If a place is subject to currents coming from polar areas, therefore colder, it will have different characteristics from one subject to equatorial currents. In reality, the warmer currents have a greater impact since the cold ones flow at greater depths due to their greater weight.

Mountain ranges

As we have seen so far, the circulation of atmospheric and marine currents has a great impact on the climatic factors for a city. The large mountain systems play a fundamental role in this circulation as they can block the circulation of a current, warm or cold, in one direction or the other. In Italy an example are the Alps that protect the peninsula from the cold currents coming from the continent and prevent the warmer currents from the Mediterranean from going to the Nordic countries. Another influence that a mountain can bring is due to the creation of particular air currents such as the föhn that invests Italy from the Alps towards the south.

Relief and topographic exposure

As the altitude increases, the layer of air above decreases in both thickness and density, resulting in a decrease in pressure. Consequently, the average temperature decreases with altitude, although not linearly, despite the increase in the amount of insolation that such a place receives.

Other factors

Three other factors can be considered to influence the climatic conditions, which will not be taken into careful consideration in the following study. The first is the characteristics of the soil, since for example more humid soils have a greater thermal capacity, heating and cooling more quickly than dry soils. A second factor is vegetation, since it ensures a certain minimum humidity and attenuates thermal extremes. In the study of temperatures, the data collection takes place in all cases in meteorological stations located near airports and it is therefore obvious to conclude that vegetation near take-off and landing strips is not frequent, unlike large expanses of concrete. With this consideration in mind, the influence of vegetation, although relevant in practice, is not to be taken into consideration in this specific case. A final factor, which is connected to what has just been said, is the influence of anthropomorphic activities. Deforestation, the draining of marshy areas, more or less polluted atmosphere and large artificial basins are just some of the works that human has implemented that have led to a modification of the climate of the place.

2.3.3 Influence of the peninsula's morphology on the climate

The Italian territory is characterized by a great climatic variability considering its limited territorial extension. This variability is due not only to the geographical position of the peninsula but also to its complex morphology. Elements such as mountain ranges, plains and their conformation, proximity to seas and lakes as well as the presence of hydrographic basins play a fundamental role in influencing the climatic characteristics for temperatures, precipitation, humidity and wind speed. One of the morphological elements undoubtedly most relevant for the Italian peninsula is the mountain range of the Alps that extends along the northern border of the country. The Alps act as a natural barrier against the colder currents coming from the north, ensuring that at the same latitude, Italian cities are slightly less cold than other European cities. However, in the Alpine valleys, significant thermal inversion phenomena are encountered, especially in the winter period. The second great Italian morphological element is the Po Valley, a large depression surrounded by the Alps to the north and the Apennines to the south. The Po Valley, due to its characteristics, is subject to atmospheric stagnation phenomena, significantly influencing the local climate. During the winter months, the poor ventilation of the plain favors the accumulation of humid and cold air, consequently leading to the formation of persistent fog and lower minimum temperatures. The lack of recirculation also influences the summer seasons, amplifying, for example, the effect of the heat. This lack of appropriate recirculation also leads to other problems such as, for example, the accumulation of fine dust due to the major industrial cities of the country that arise precisely in this area. The Apennines, which divide Italy in half from north to south, represent a further morphological element capable of influencing the climate of the peninsula. Their presence affects not only the variability of temperatures between the Tyrrhenian and Adriatic sides but also the distribution of precipitation. The western side, more exposed to Atlantic disturbances, receives greater precipitation while the eastern side is more influenced by the presence of the Balkans. Significant differences in average temperatures can be observed between the two coasts, with the Tyrrhenian side characterized by a milder and more humid climate, while the Adriatic side has colder winters and greater annual temperature variations. The greatest factor influencing the Italian climate is undoubtedly the Mediterranean Sea, which plays a very important role in mitigating the temperatures of coastal cities by reducing temperature variations as the large mass of water absorbs heat during the summer months and then gradually releases it in the colder periods, mitigating winter temperatures. This effect is particularly evident as coastal cities see higher minimum temperatures than those inland in the same region. A final morphological element to take into consideration are the lake basins, which exert an effect very similar to that of the seas but on a local scale. The large lakes in the Prealps such as Lake Garda or Lake Maggiore contribute to the slightly more temperate climate of the surrounding areas.

2.4 Climate Zones of Italy: An Overview of Temperature Variations

2.4.1 Pinna and Köppen climate classification

One of the most widespread and well-known classification systems is the one created by the German climatologist Wladimir Köppen at the beginning of the 20th century. His classification aims to provide a clear and systematic representation of the different climatic areas of the planet using measurable factors such as temperatures and precipitation. The classification involves the division of the world's climates into five large categories represented by the capital letters from A to E (A: tropical, B: arid, C: temperate, D: continental and E: polar). These macro-categories are in turn subdivided according to precipitation and temperature. The analysis to determine the classification is based on quantitative criteria, such as the average temperature of the coldest month and that of the warmest month, as well as specific precipitation thresholds, which makes it a useful tool for understanding global climate distribution. However, since it was developed on a global scale, the Köppen model does not always accurately reflect the local climate characteristics of some regions, which is why it has been subject to local adaptations that take into account the different characteristics of the place. The classification is presented in Figure 2.3. Mario Pinna, one of the most prominent figures in Italian climatology, reinterpreted

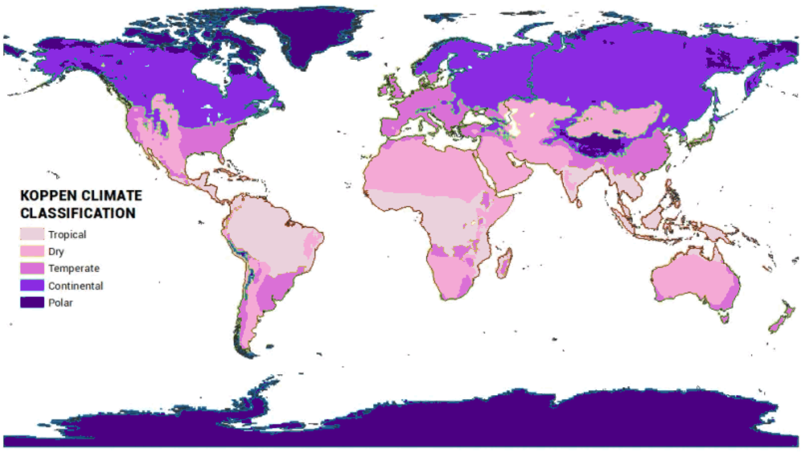


Figure 2.3: Division of the world into the 5 Köppen categories.
Source: `https://earthhow.com/koppen-climate-classification/`

Köppen's classification to adapt it to the characteristics of the Italian peninsula in [17]. At the heart of his reinterpretation is a different distinction between the subcategories of temperate climates (type C) following a redefinition of the criteria for defining the arid month. According to Köppen, a Csa (Mediterranean with hot summer) or Csb (Mediterranean with temperate summer) climate is characterized by a dry summer, defined by the criterion according to which the driest month of the summer must have precipitation lower than a certain value linked to the average annual temperature. However, Pinna believed that this criterion did not adequately apply to Italy, where the Mediterranean climate has a complex behavior due to the complex morphology with abundant local variations due to the different characteristics of the individual territory. For this reason, he proposed a modification establishing that a month was considered arid if precipitation was lower than a certain threshold, regardless of the average annual temperature. Thanks to this modification, Pinna redefined some Italian climatic areas, identifying as Csa and Csb zones that, according to the original Köppen classification, would have been included in the Cfa (humid temperate without dry season) or Cfb (oceanic) climates. This adaptation has allowed to better represent the Italian climatic variety, highlighting how the Mediterranean climate also influences areas of central and northern Italy, which in the standard Köppen classification would have been considered more similar to those of central Europe. The classification of the Italian climate zones is reported in the Figure 2.4.

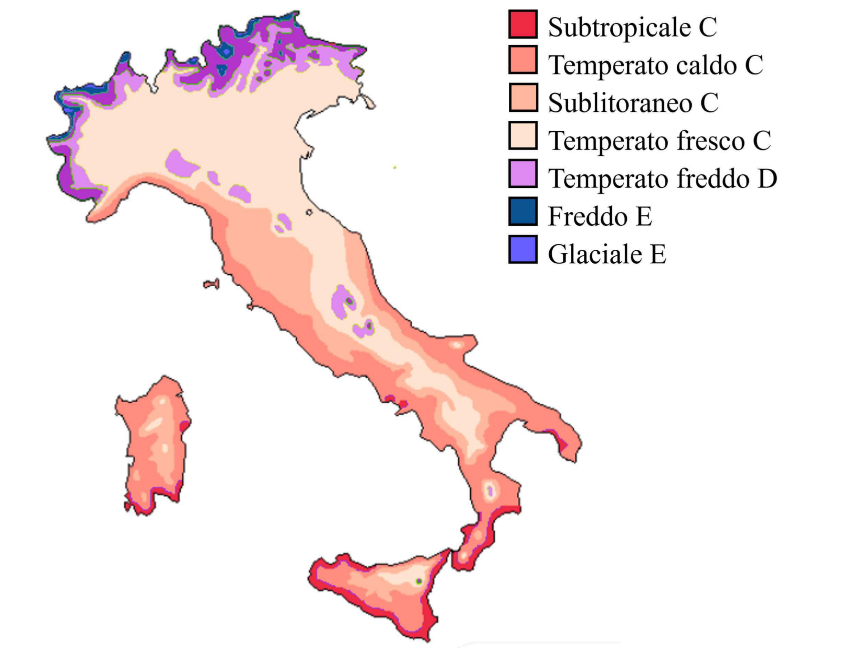


Figure 2.4: Division of Italy into the Pinna climate zones.
Source:`https://it.wikipedia.org/wiki/File:Italia_Clima_Pinna.png`

2.4.2 Simplified climate classification

Although the classification proposed by Pinna is the result of a rigorous scientific study based on Koppen's classifications for the Italian climate, for the purposes of the analysis that will be proposed, it was preferred to adopt a more simplified climatic subdivision that takes into account easily recognizable geographical criteria. The following classification takes into account the main distinctions identified by Pinna, but reorganizes them into six large climatic areas: Mediterranean, Alpine, Apennine, Adriatic, Po Valley and Tyrrhenian. This approach allows not only to simplify the subdivision of Italy into climatic zones but also to more clearly delineate the predominant climatic influences in the different Italian regions, simplifying the reading and interpretation of climatic variations. In this way, a link with Pinna's rigorous approach is maintained, but a more immediate reading of Italian climatic dynamics is favored. The classification proposed takes its cue from the climatic classification of Italian municipalities introduced by the Presidential Decree n. 412 of 26 August 1993 in Article 2 in which the Italian territory is divided into six climatic zones, named from A to F based on the average daily external temperature. The simplified version of the classification is proposed in Figure 2.5.

In the Alps and Prealps the dominant climate is Alpine with increasingly colder temperatures as the altitude increases. The Italian side of the Alps, being exposed to the south, tends to enjoy slightly milder conditions than the eastern and northern sides as the mountain range itself acts as a barrier to cold winds coming from the north. It is possible to notice a certain difference between the temperatures of the eastern and central Alps compared to the western ones which are slightly less cold due to the greater proximity to the Atlantic Ocean and the greater distance from central Europe. Precipitation in this area is very abundant especially in the intermediate seasons. This area of Italy is characterized by cold and harsh winters with temperatures that fall below zero Celsius accompanied by abundant precipitation, including snow. Summers are shorter and cooler with average temperatures around 20 degrees Celsius. Continuing south of the Alps, the Po Valley is characterized by a continental climate that has large daily and seasonal temperature variations. The morphology of this territory has a huge impact on the climate of the area. Consider how it is a region closed between the two mountain chains of the Alps and the Apennines that hosts within it a network of rivers and canals that favor the increase in humidity. The western and central portion tends to record colder temperatures than the eastern area more subject and open to the influences of the Adriatic Sea. Characterizing the Po Valley are long and cold winters with temperatures that can drop below zero Celsius and persistent fog. Summers, on the other hand, are hot and muggy with maximum temperatures that easily reach 35 degrees Celsius. Note in particular the abundant rainfall in the intermediate seasons. Proceeding south the climate tends progressively to be milder thanks to the greater influence of the Mediterranean Sea and the anticyclones coming from Africa. A major impact on the climate of central and southern Italy is provided by the presence of the Apennine mountain range that divides the Italian peninsula into two parts, the Tyrrhenian and the Adriatic. On the Apennines (Apennine Climate) winters are cold and accompanied by abundant snowfall while summers are moderately hot due to the greater influence of the Mediterranean Sea compared to the Alps. The Tyrrhenian side, more exposed to Atlantic disturbances that do not reach the Adriatic side in the same way thanks to the barrier of the Apennines, is characterized by more abundant precipitation as well as generally more temperate winters with temperatures that rarely fall below 5 degrees Celsius and hot summers but mitigated by the sea breeze. Furthermore, the Tyrrhenian Sea, larger and deeper than the Adriatic Sea, exerts a greater mitigating effect so much so that cities that are not located directly on the coast but slightly inland are greatly affected by the maritime influence. An example is the city of Rome which, despite being 20km from the sea, enjoys a climate typical of coastal cities. On the Adriatic side, however, the climate is more affected by cold currents coming from the Balkans which are shielded by the Apennines and thus do not reach the Tyrrhenian side, therefore winter temperatures tend to be more rigid while summer temperatures are very similar. To conclude, the two main Italian islands, Sicily and Sardinia, have a climate strongly influenced by latitude rather than by the mitigating effect of the sea. Sicily, being further south, experiences very short and mild winters, while summer is often dominated by heat waves coming from North Africa. Sardinia, however, is more affected by Atlantic disturbances and the Balearic depression, recording greater climatic variability than Sicily. In both islands, autumn is the rainiest season, followed by winter, while spring is generally mild and summer is hot and dry.

Figure 2.5: Division of Italy into climatic zones.
Source: `https://www.gmpe.it/sites/default/files/meteorologia/italia.png`

2.4.3 Selected cities and their connection with the climate zones of Italy

The objective taken into account when selecting the Italian cities on which to base the following analysis was to try not only to represent the entire peninsula but also the different climate zones. Each city was chosen taking into account its geographical position, meteorological characteristics, and the influence of the surrounding territory to best represent Italy. The selected cities cover all the different climate zones present in Italy in order to offer the broadest possible overview. The choice of only cities and the lack of remote areas, and perhaps with more particular and interesting climatic conditions, is due to a practical question. Historical series with reliable measurements are not easy to find for remote places of the peninsula while for large urban centers that as such have an airport, have at their disposal specific meteorological stations from which it is possible to obtain with precision the data used later. The first city, proceeding in order from west to east and from north to south, is Turin. Located in the western Po Valley, it has a continental climate (or Po Valley as defined in the previous paragraph) with harsh winters and hot and muggy summers. It is impossible not to consider the presence of the Alps that dominate the horizon to the north and west of the city, creating a natural barrier against the cold winds coming from northern Europe, influencing the climate of Turin. The second city considered in this study is Bolzano. It is the only representative of the Alpine climate as it is located in a basin of the southern Alps. The South Tyrolean city has typical characteristics of the Alpine climate with not too hot summers and cold and snowy winters. Moving for the first time to the coast, we find Venice on the shores of the northern Adriatic Sea. The Venetian city influenced by the presence of the sea has an Adriatic-type climate that allows it to mitigate temperatures with mild winters and hot and humid summers. Bologna, although also in the Po Valley together with Turin and being part of the same climate zone, is more affected by the maritime influence, thanks to its proximity to the Adriatic Sea, which gives it a climate with hot but less extreme summers. Continuing south in Central Italy, Viterbo falls within the Apennine area, surrounded by the Apennine hills to the east and the volcanic sands of the Cimini area to the north, and has hot summers but with colder winters than the coastal areas. The Tyrrhenian area is represented, however, exclusively by Rome with its climate characterized by hot and dry summers and mild winters. On the other side of the Apennines, however, Ancona is also located on the Adriatic coast. Its climate is undoubtedly influenced by the sea, with hot summers, although less than the Adriatic coasts further south, and is another example of the Adriatic climate. In the South, Bari and Palermo represent the Mediterranean areas, with a hot and dry climate in summer and mild winters. However, the position of Palermo, in the western part of Sicily, makes it particularly influenced by the thermal dynamics of the Mediterranean Sea, resulting in higher summer temperatures than Bari, which is located further east and more exposed to the cold winds of the Balkans.

2.5 From raw data to clean data: Preprocessing Steps

2.5.1 Data collection and aggregation

The temperature data used for the analysis, as already mentioned, are taken from the website *ilmeteo.it*. These data were available in monthly files. For each city, the data relating to the last twenty years, 2005-2024, were downloaded, so for each city there was a total of 240 different files. Each of these files contained not only the temperature readings specific to the selected city but also a series of other meteorological parameters that are reported in the Table 2.1. However, for the purposes of our analysis, we focused exclusively on temperature data, in particular Tmin, Tmax and Tmedia. These parameters represent key indicators for the analysis of climate variations and thermal trends.

Column	**Description**	**Data Type**	**Unit of Measure**
LOCALITA	Location	String	
DATA	Date (dd/mm/yyyy)	Date	
TMEDIA	Average temperature	Integer	°C
TMIN	Minimum temperature	Integer	°C
TMAX	Maximum temperature	Integer	°C
PUNTORUGIADA	Dew point	Integer	°C
UMIDITA	Humidity	Integer	%
VISIBILITA	Visibility	Integer	km
VENTOMEDIA	Average wind speed	Integer	km/h
VENTOMAX	Maximum wind speed	Integer	km/h
RAFFICA	Wind gust	Integer	km/h
PRESSIONESLM	Sea level pressure	Integer	mb
PRESSIONEMEDIA	Average pressure	Integer	mb
PIOGGIA	Rainfall	Integer	mm
FENOMENI	Phenomena	String	

Table 2.1: Meteorological dataset columns and data types.

The first step was therefore to merge all the monthly files into a single dataset for each city, thus creating a continuous and complete historical series. To support this operation, the Python code used to merge the files and build the overall dataset is reported in the Appendix.

2.5.2 Cleaning and conversion of numerical columns

A first fundamental phase before proceeding with an accurate analysis of the data concerns the conversion of the data that in the Table 2.1 are reported as numeric in a numeric format suitable for statistical and mathematical manipulation. The raw data, in fact, presented textual elements or symbols that prevented a direct interpretation of the values as real numbers, making preliminary cleaning work necessary. For this reason, a cycle was written that iterates on all the columns involved, including those fundamental for the subsequent analysis such as minimum, maximum and average temperatures. For each column, the values were all converted into strings in order to be able to work uniformly on the entire column and to be able to use the text manipulation tools. Subsequently, a systematic cleaning of the data was carried out:

- the spaces inside the strings were removed,
- all non-numeric characters, therefore all except digits, decimal point, comma and minus sign, were eliminated,
- commas, used in some cases as decimal separators, were replaced with a period to comply with Python's numeric syntax.

The last step was to convert the data into numerical values, paying particular attention to non-interpretable values that were set as NaN. For completeness, the code used in this phase has been reported in the Appendix.

2.5.3 Feature engineering

In order to work more easily with the dataset, new variables have been created from the original data. In particular, from the date of each data-point, the month and year variables have been created to facilitate the analysis of monthly and seasonal trends as well as annual ones. The final variable introduced, called the trend, is a progressive number that starts from 1 and continues until the end of the data set. This new variable will be very useful later when we want to define a model for the data at our disposal, as this variable will be a good support for highlighting any presence of long-term changes in the data attributable, for example, to climate change.

2.5.4 Managing missing data and outliers

During a first exploratory phase of the data, both NaN values and outliers were found. In the first case, identification is simple using the *np.isnan()* command provided by the Python NumPy library. In the case of outliers, their identification was slightly more complex. These steps are essential to ensure effective and reliable data analysis. In the Table 2.2 you can see the number and distribution of NaNs and outliers for different cities.

$N°$	Turin	Bolzano	Venice	Bologna	Viterbo	Rome	Ancona	Bari	Palermo
Data points	7275	7275	7275	7275	7275	7275	7275	7275	7275
NaN	37	39	34	27	126	32	26	45	36
Tmin outlier	2	1	4	2	3	2	3	2	27
Tmax outlier	1	1	4	1	2	2	1	1	1
Tmean outlier	2	1	2	2	2	2	1	2	2

Table 2.2: Before preprocessing.

Handling NaN values

In data preprocessing, handling NaN (Not a Number) values is a key task to ensure the integrity and consistency of the dataset. In many cases, the entire row for a given day had NaN values for maximum, minimum, and mean temperatures, which could compromise the quality of the analysis. To address this issue, it was adopted a two-step approach.

First, for the missing values for maximum and minimum temperatures, we chose to replace them with the average of the values observed in the 7 days before and 7 days after the date when the value was missing. This method proved to be particularly useful as it takes into account seasonal and weekly variations in temperature, allowing for more realistic estimates of the missing values without introducing significant bias.

Second, for the daily mean temperature, it was applied a similar methodology: the NaN values were replaced with the average of the two values calculated in the previous step, i.e. the maximum and minimum temperatures for each day. This approach allowed to obtain a plausible value for the mean temperature even when the two extremes (maximum or minimum) were absent.

These substitution methods were chosen for their simplicity and effectiveness in preserving the underlying trends and patterns of the data. The specific code used to implement these operations will be provided in the Appendix.

Identify and resolve outliers

In the Figure 2.6, it is possible to see two examples of outliers. In the plot on the left, we have the mean, minimum, and maximum temperatures for the city of Palermo in September 2011. It can be observed that all three quantities were erroneously set to $0°C$. The dashed line represents the values for the specific date after applying the method described for outlier resolution. On the right, in the plot of temperatures for Turin in May 2010, it is evident that the erroneous temperatures are only the minimum and mean, marked as $0°C$ and $3°C$, respectively. By analyzing the plots for the different cities on the two dates where the outliers are present in Figure 2.6 it was possible to find that in all cases the data of 09/14/2011 and 24/05/2010 appear to be somehow incorrect. To identify the outliers, two approaches were followed. The first was a numerical method with which possible outliers were identified by looking at the distance between the previous

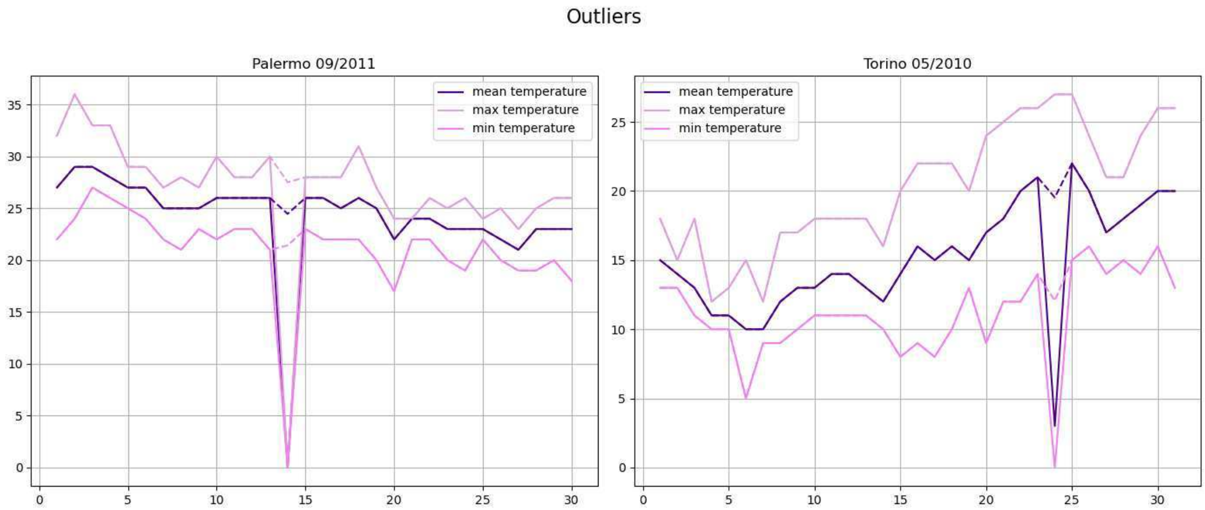


Figure 2.6: Examples of outliers for Turin and Palermo.

and subsequent measurements. If these differences were greater than a threshold set at 3 degrees Celsius, then that specific data point was analyzed in more detail to understand whether or not it was an outlier. This first identification method was excellent in reporting the outliers even if it presented many false positives. For the second method, it was decided to use a more graphical approach. Serial graphs were performed both of the single month for each year and of that specific month over the years in order to verify anomalies of a specific month compared to the monthly average over the twenty years analyzed. With this last approach, it was possible to find that the year 2010 was significantly colder in the winter months than in previous years as can be seen from Table 2.3.

	1	2	3	4	5	6	7	8	9	10	11	12
2010	1.258	3.926	8.452	13.567	17.744	22.367	26.323	23.774	18.767	12.677	9.433	1.806
Monthly mean	3.484	5.704	9.753	14.022	18.404	23.316	26.115	25.259	20.606	15.291	9.444	4.307

Table 2.3: Monthly averages for 2010 and monthly averages for the last 20 years.

2.6 Data analysis

In this section the main objective is to have some insight on the temperature.
In Table 2.4 it is possible to observe the values relating to the average temperatures, with the minimum, maximum and variance for the different cities. It is immediately possible to notice how there are significant variations between one season and another, suggesting that in the subsequent analyses the division into seasons if not months will be fundamental to better understand the dynamics of the temperatures. In detail we observe:

- The average temperatures show strong variations between the different cities under examination. For example, Palermo shows very high average temperatures with the winter average exceeding $13°C$ while cities in the north of the Italian peninsula such as Bolzano and Turin do not exceed the seasonal average of $4°C$ in winter. The average temperatures then tend to become uniform both in spring with all the values hovering around $14-15°C$ and in summer with the only exception of Turin which is slightly cooler.

- The seasonal variances provide us with fundamental indications on the dispersion of the data. We can observe how inland cities such as Bologna and Bolzano tend to have a fairly high variance especially in autumn and spring. On the other hand, cities such as Palermo and Bari, which overlook the sea, show much lower variances.

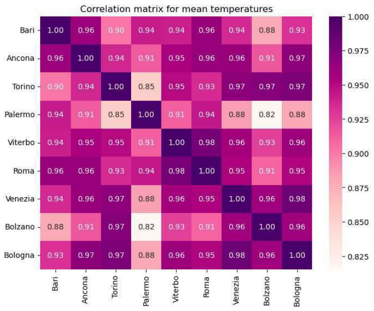
- Minimum and maximum temperatures highlight how seasonal fluctuations can vary considerably. For example, the city of Palermo has relatively high winter minimum temperatures (around $5°C$), while cities such as Bolzano and Turin can record minimums below zero, with extreme values reaching even $-10°C$ in winter. Similarly, summer maximums are much higher in Palermo compared to other northern cities such as Turin, where maximum temperatures are recorded below $30°C$. These extreme values suggest that seasonality significantly affects temperatures, making the division into seasons useful for a more precise analysis.

	Fall				**Spring**				**Summer**				**Winter**			
City	Mean	Var	Min	Max	Mean	Var	Min	Max	Mean	Var	Min	Max	Mean	Var	Min	Max
Bologna	15.12	30.75	1.0	28.0	14.06	23.24	-4.1	28.0	24.91	9.77	12.9	33.0	4.46	10.18	-8.0	14.0
Bari	17.66	19.17	5.0	29.0	14.84	16.82	1.4	27.0	25.17	8.76	13.7	36.0	9.24	8.18	-1.0	18.0
Bolzano	13.48	37.51	-3.3	27.0	14.85	23.45	-3.5	28.0	24.27	10.82	13.0	33.0	3.00	11.13	-7.0	14.0
Viterbo	16.92	24.74	2.0	29.0	14.67	20.52	-1.3	28.4	25.81	12.05	12.2	35.0	7.83	9.47	-4.0	17.0
Roma	17.35	20.86	3.0	30.0	14.57	15.63	2.3	27.0	25.04	8.40	14.2	32.0	8.54	9.53	-4.0	18.0
Palermo	21.03	12.98	11.0	32.0	16.49	9.99	7.6	32.0	25.57	6.16	16.5	38.6	13.21	5.75	5.0	23.0
Ancona	15.99	21.55	2.0	29.0	13.83	18.78	-1.0	27.0	24.11	8.16	13.0	33.0	6.89	10.03	-4.0	22.0
Venezia	14.91	26.75	1.0	27.0	13.71	21.64	-2.1	27.0	24.06	7.57	14.0	31.0	4.72	8.43	-7.0	13.0
Torino	13.63	28.80	0.0	26.0	13.14	20.93	-3.9	25.4	22.92	8.45	13.0	30.0	3.67	9.41	-10.0	17.0

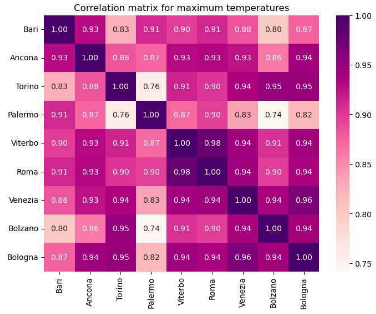
Table 2.4: Temperature statistics for different cities by season over the last 20 years.

2.6.1 Correlations between cities

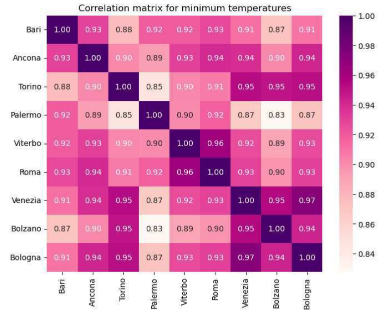
A first step in the analysis of temperatures is the calculation of the correlation matrices for the three variables analyzed. The three matrices reported in the Figure 2.7 provide only a general overview of the relationships between the different cities but are not very useful for understanding the different temperature dynamics as they do not allow to capture the seasonal variations specific to each place. Therefore, for a more precise understanding of the correlations, it is necessary to proceed with the seasonal analysis of temperatures, as described below. The Figure 2.8 shows the temperature correlation matrices for the different cities examined, dividing them into seasons. For the division into seasons, it



(a) Mean temperature.



(b) Max temperature.



(c) Min temperature.

Figure 2.7: Correlation matrices for mean, max and min temepratures.

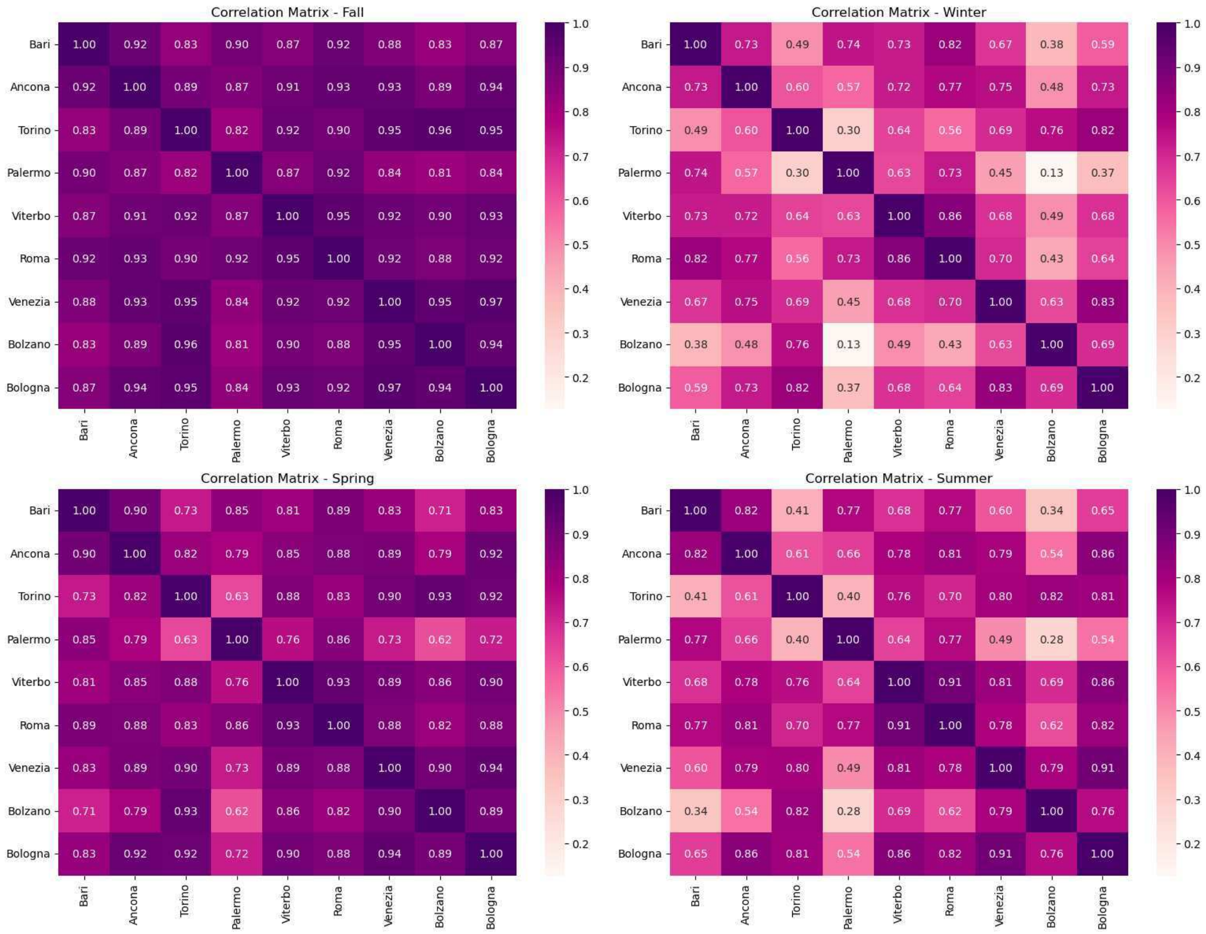


Figure 2.8: Seasonal correlation matrices.

was preferred to rely on whole months, rather than following the traditional astronomical division based on equinoxes and solstices. This choice, quite common in the climatological field, facilitates the management and processing of data, allowing more coherent and uniform comparisons between seasonal periods. In detail, winter was defined as the quarter December-January-February, spring as March-April-May, summer as June-July-August and autumn as September-October-November. This approach is also consistent with the conventions adopted in the field of weather derivatives, where it is preferred to use months rather than astronomical seasons, in order to guarantee a clear, regular temporal basis that can be easily integrated into quantitative models. From the seasonal correlation matrices of temperatures for the nine Italian cities it is possible to better understand the various temperature dynamics mentioned above. During fall very high correlations are observed between almost all cities, with all values above 0.80. The cities of the Center-North show very high correlations among themselves, but even geographically more distant locations such as Palermo show quite high values, suggesting that in this season the temperature variations are due to the same factors as they follow a fairly similar trend throughout the country. In winter the situation is very different because the correlations become much weaker. An example is Bolzano which shows particularly low correlations with the other cities, reflecting its more rigid Alpine climate, or Palermo which clearly stands out from the rest of the cities in line with the mild Mediterranean climate of Sicily. However, cities such as Rome, Ancona and Bologna maintain higher correlations among themselves, given their temperate climate and a smaller temperature range. In winter, therefore, regional differences are accentuated: the different exposure to cold and mild air masses, combined with the orography of the territory, causes temperatures to vary in a less synchronized way on a national scale. In spring, as happened in the other intermediate season, it is possible to observe a general increase in correlations, especially between the cities of the Center-North. Turin, Venice, Bologna, Ancona, Rome and Viterbo show very high values, and Bolzano also returns to present high correlations, a sign that the increase in daylight hours and the attenuation of seasonal thermal differences bring temperatures back to a more homogeneous behavior. Palermo continues to maintain slightly lower values than the rest of the country. During summer, the matrix shows a less uniform trend as in winter. Bolzano shows low values with the cities of the South such as Palermo and Bari, but maintains good correlations with Venice and Bologna. Turin also tends to have a behavior similar to that of the city in Alto Adige. A certain internal coherence can be noted between cities in the Centre such as Rome, Viterbo and Bologna which show very high values among themselves, a sign that the central area of Italy responds in a more synchronized way to summer temperature trends. In general, summer brings with it local phenomena, such as heat waves, sea breezes and orographic influence, which determine inhomogeneous thermal behaviors even between relatively close cities. The analysis therefore shows that autumn and spring are the seasons in which temperatures tend to have a more similar behavior on a national scale, while winter and summer highlight greater local specification, linked to climatic and geographical factors such as latitude, altitude and distance from the sea. By analyzing the Figure 2.9 representing the evolution of the annual seasonal averages, it is possible to reinforce what emerged from the correlation matrices. During the winter, strong differences are observed between the cities with reference to the lines of different colors that refer to each of the cities that are visibly more scattered, with Palermo stably confirming itself as the warmest city and Bolzano the cold-

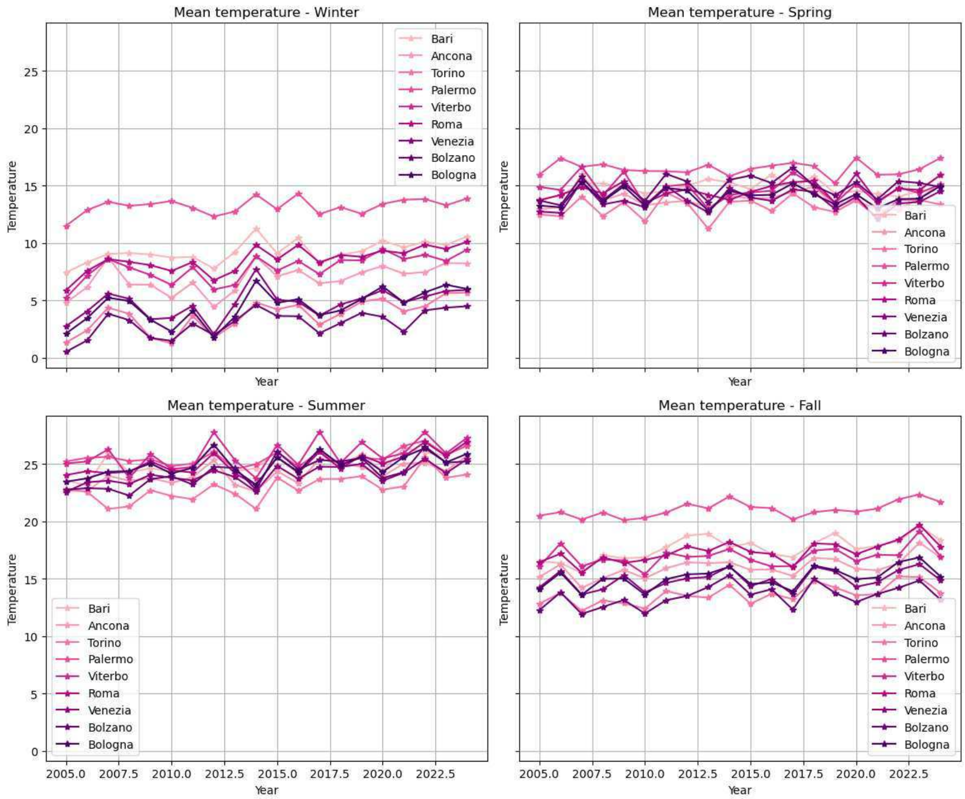


Figure 2.9: Annual mean temperature divided by season.

est. This divergence, also visible in the dispersion of the values in the graph, explains the low winter correlations observed between geographically distant cities. On the contrary, in spring and autumn the temperature lines appear much closer together, suggesting a more uniform thermal behavior between the cities as already observed previously with the analysis of the values of the correlation matrix for the spring and autumn seasons. The summer, while maintaining high temperatures in all cities, presents more marked fluctuations and detachments, especially between Palermo and the cities of the North, which explain the variability of the summer correlations. Furthermore, in all seasons there is a general trend towards an increase in temperatures, particularly marked in the summer months, confirming a progressive warming that will have to be taken into account when defining a model for temperatures.

The analysis of the correlation matrices allowed us to highlight patterns of similarity and discontinuity that vary significantly with the change of season. During the winter, for example, the correlations between cities in Northern Italy, such as Turin, Milan and Bologna, are high, confirming a homogeneous thermal behavior in this area. On the contrary, southern cities such as Palermo or Bari show weaker correlations with the North, reflecting a different seasonal climate response. To better understand the spatial distribution of temperature, we decided to use a geostatistical technique, ordinary Kriging. Kriging allows us to estimate the values in unobserved points by exploiting the spatial autocorrelation of the analyzed variable, modeled through a variogram. This approach returns continuous interpolated surfaces, offering a visual picture of the distribution of seasonal temperatures, and allowing us to capture climatic transitions that are not immediately evident if we only observe the point data.

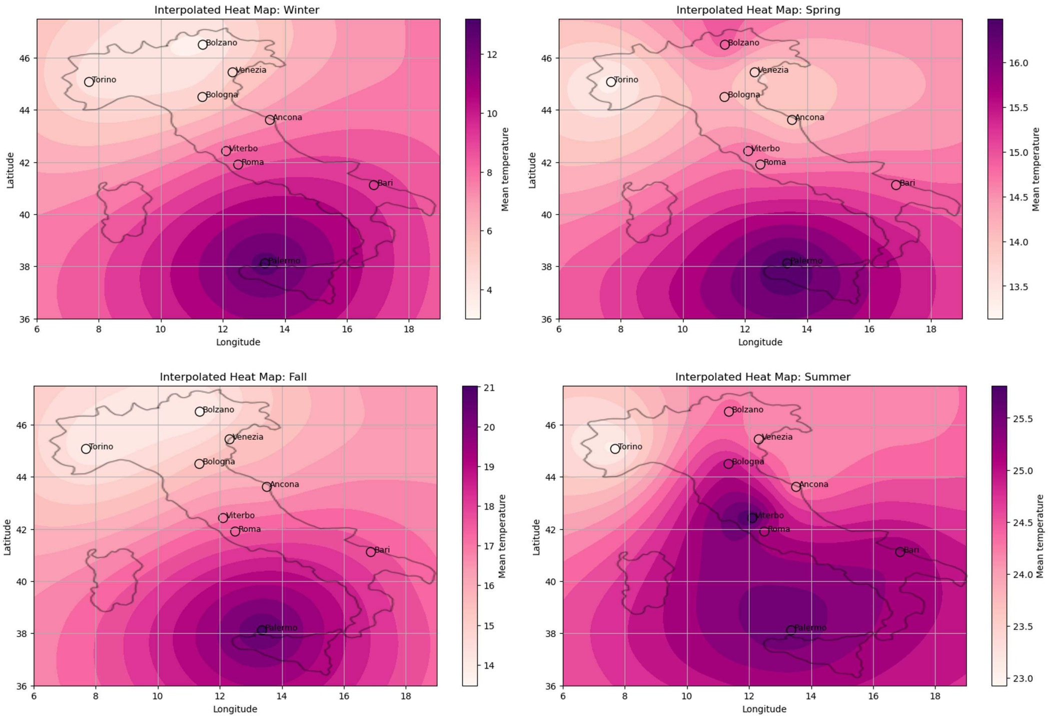


Figure 2.10: Kriging interpolation of seasonal temperatures.

The maps obtained in Figure 2.10 clearly show how the seasonal mean temperature distribution follows a well-defined latitudinal gradient, especially in winter and autumn, where the Northern regions, such as Bolzano and Turin, are significantly colder than the southern cities, with Palermo representing a constant positive thermal outlier. In these periods, the central areas are located in an intermediate range, as also confirmed by the more moderate correlations with both geographical extremes. In the summer period, the gradient attenuates, as also highlighted by the correlation matrices that show a general increase in thermal homogeneity between cities. The interpolated maps confirm this behavior: high temperatures are distributed more uniformly along the entire peninsula, with reduced differences between North and South. However, there are slight local differences influenced by topographic factors and continentality, as observed in the case of Turin, which is slightly cooler than the surrounding context. During spring, an interesting thermal transition is observed: Palermo maintains significantly higher values than other cities, but the areas of the Center-North begin to approach the values of the South, anticipating the summer uniformity. It should also be considered that the southern cities located along the coast tend to warm up more slowly in spring than those inland.

2.6.2 Trend analysis

Undoubtedly, one of the most studied and discussed topics of the modern era is climate change. The need to understand its causes and predict its effects has made it the center of scientific, political, and economic debates due to the numerous implications that climate change can have at the environmental, social, and economic level. According to the definition provided by the World Meteorological Organization (WMO), climate change is "the term used to describe changes in the state of the climate that can be identified by changes in the average and/or the variability of its properties and that persists for an extended period, typically decades or longer." [20]. The data recently published by the WMO in [20] clearly highlight the extent of the phenomenon: 2024 was the warmest year ever recorded in the 175 years of instrumental observations, with a global mean temperature close to the Earth's surface of $1{,}55 \pm 0{,}13\,^{\circ}\mathrm{C}$ above the average value of the period 1850-1900. Added to this is the data on the global mean sea level, which in 2024 reached a new historical maximum since the beginning of satellite observations in 1993, with an average rise of 113 mm compared to the levels taken as a reference. Considering that climate change is a concrete and scientifically documented reality, the definition of a linear trend in the historical data for the temperatures considered in this work appears as a consideration as natural as it is necessary. The observed trend of increasing average temperatures on a global scale suggests that the thermal behavior of the atmospheric system can no longer be described as stationary. In this context, neglecting the presence of a trend would mean ignoring a fundamental component of contemporary climate dynamics. On the contrary, recognizing and quantifying its existence allows us to build models that are more in line with reality, capable of capturing the actual long-term direction in which the climate is evolving. The inclusion of the linear trend, despite its simplicity, therefore represents a first modeling response consistent with the evidence of global warming and constitutes a solid starting point for any future refinements. The aim of the following assessment is to understand whether there is a significant thermal trend in the data, in order to decide whether or not to include a trend component within the climate model that is going to be developed later. In a first exploratory approach, the average annual temperatures for each city were calculated and reported graphically over the years. In Figure 2.11 it is possible to see a clear hint of a positive trend, consistent with what is highlighted in the main reports on climate change. To confirm our intuition it is necessary to continue with a more statistical analysis of the trend. To do this it was decided to follow two approaches. In the first, daily temperatures were analyzed without seasonal distinction and aggregated only annually, while in the second, it was decided to group them also taking into account the meteorological season — Winter (December–February), Spring (March–May), Summer (June–August) and Autumn (September–November) — to better understand the underlying dynamics of temperature evolution that would be

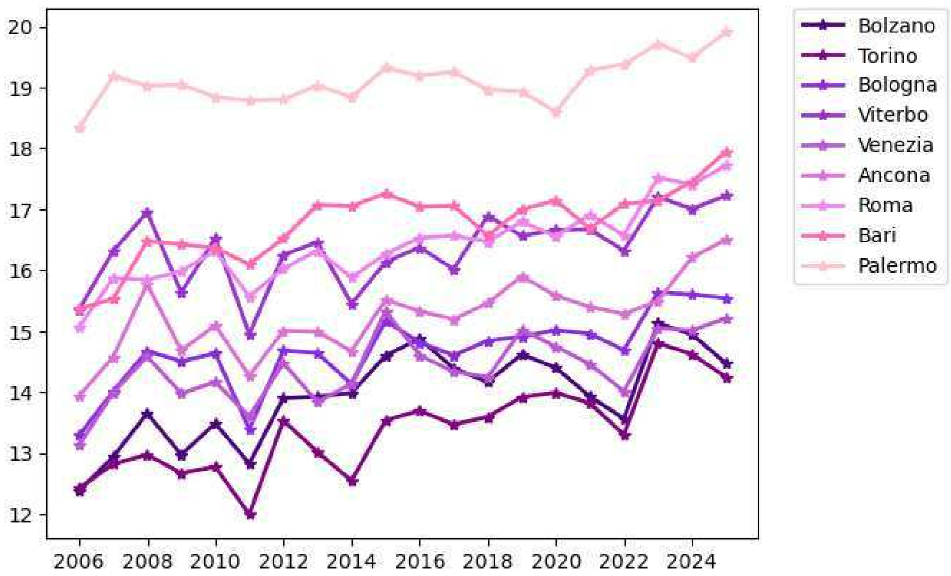


Figure 2.11: Annual mean temperature.

masked when considering only annual data. On each city-season time series, as well as on the annual one, a linear regression analysis was applied in order to estimate the temperature trend over time, expressed as an angular coefficient in °C/year. This allowed to quantify both the direction and the intensity of the thermal change observed in the period considered. The statistical significance of each trend was assessed through the p-value, adopting a conventional threshold of 0.05. The results of seasonal and annual

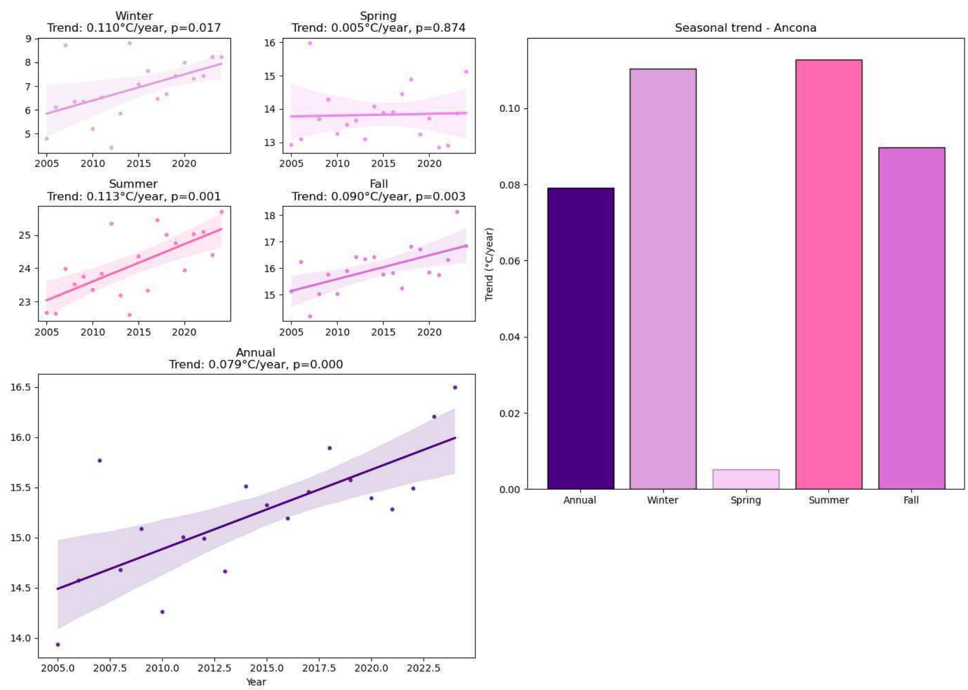


Figure 2.12: Trend analysis for the city of Ancona, based on linear regressions. The istogram displays the obtained angular coefficients.

regressions for the city of Ancona are reported in Figure 2.12. As it can be seen, the angular coefficient of the regression with the annual mean temperatures has a positive value ($0.079°C/year$) and is statistically significant. Looking only at the annual data, an interesting detail would have been missed: the average temperatures tend to increase in all seasons, especially in winter and summer, while in spring they remain almost stable. The considerations just made for the city in the Marche region are also valid, with a few exceptions, for the other eight cities analyzed. As can be seen in Figure 2.13, where the transparent bars of the histogram represent the non-significant results, the angular coefficients of the seasonal regressions in spring are not significant for each city. If we take into account only the statistically significant values, it is possible to observe how, in any case, the average temperature tends to increase, especially in the winter months.

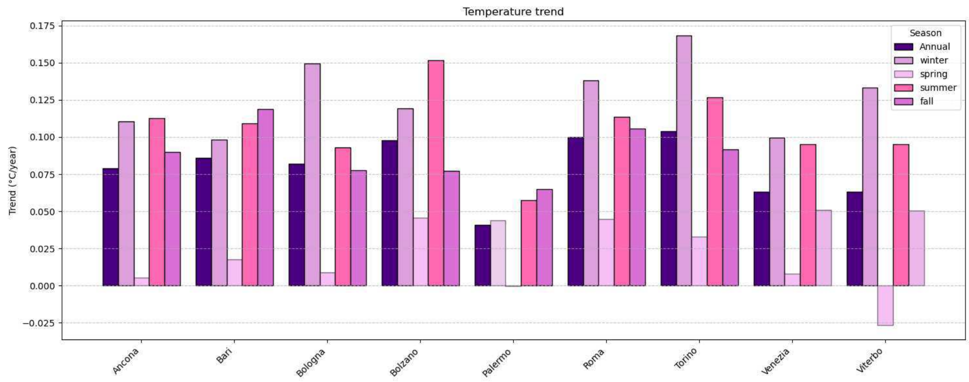


Figure 2.13: Linear regression coefficients (°C/year) for each city by season and annually. Transparent bars indicate non-significant trends ($p > 0.05$).

The regression results with their respective values and p-values are reported in the Table 2.5 where the statistically significant values have been inserted in bold. The plots showing the linear regression and average temperatures are presented in the Figure 2.14 for the annual data, while those related to seasonal data are provided in the Appendix.

2.6.3 Volatility analysis

In the study of the behavior of temperatures, a relevant factor is thermal instability. To measure this instability in the temperatures under analysis, volatility provides us with excellent indications on how much their behavior is more or less regular. To study climate variability there are different ways to estimate temperature volatility. One of the classic and most widespread approaches is the one that estimates volatility using the deviation of daily values from the mean value of the period. However, in this work we preferred to use a different estimator that uses the deviation between two consecutive measures instead of the deviation between the daily value and the mean. This choice is motivated by two main factors. First, the estimate built in this way better reflects local temperature fluctuations, that is, how much it varies from one day to another without depending on

	Winter		**Spring**		**Summer**		**Fall**	
City	value	p-value	value	p-value	value	p-value	value	p-value
Bari	**0.0981**	**0.0046**	0.0178	0.4956	**0.1092**	**0.0041**	**0.1189**	**0.0008**
Ancona	**0.1104**	**0.0168**	0.0052	0.8739	**0.1129**	**0.0010**	**0.0897**	**0.0033**
Torino	**0.1685**	**0.0004**	0.0328	0.3335	**0.1269**	**0.0011**	**0.0918**	**0.0021**
Palermo	0.0437	0.1015	-0.0004	0.9874	**0.0575**	**0.0214**	**0.0650**	**0.0060**
Viterbo	**0.1333**	**0.0013**	-0.0266	0.4931	**0.0951**	**0.0469**	0.0506	0.1209
Roma	**0.1382**	**0.0002**	0.0449	0.1369	**0.1134**	**0.0015**	**0.1059**	**0.0013**
Venezia	**0.0994**	**0.0398**	0.0082	0.8186	**0.0950**	**0.0009**	0.0509	0.1148
Bolzano	**0.1194**	**0.0042**	0.0458	0.2281	**0.1518**	**0.0003**	**0.0770**	**0.0393**
Bologna	**0.1494**	**0.0037**	0.0087	0.7791	**0.0930**	**0.0099**	**0.0777**	**0.0203**

Table 2.5: Seasonal trends of annual mean temperatures (°C/year) and related p-values for each city. Statistically significant values are in bold($p < 0.05$).

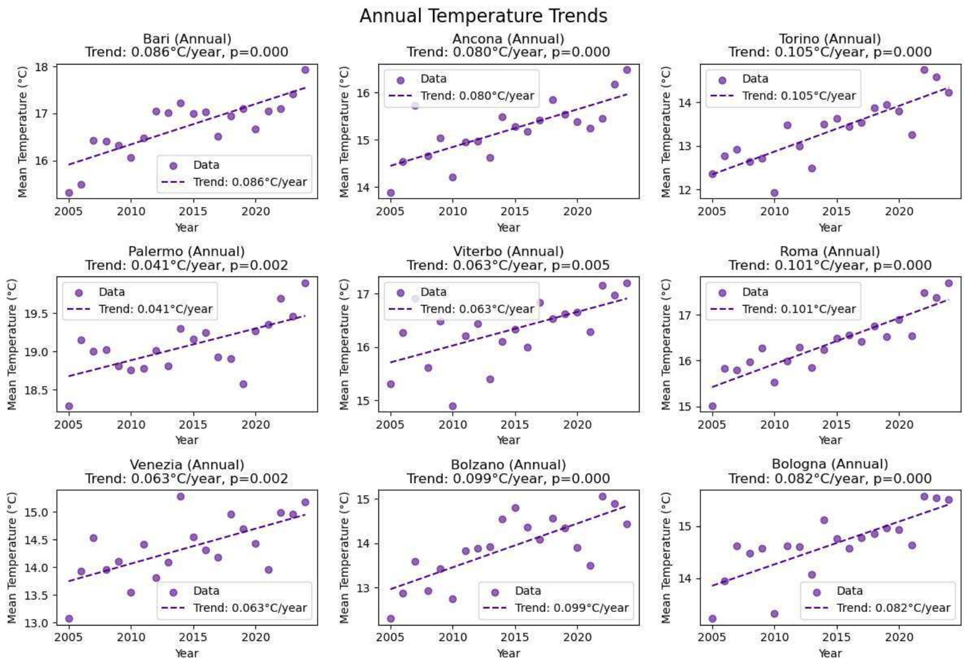


Figure 2.14: Annual linear regression coefficients (°C/year) for each city.

a global mean value that may not be representative of the chosen period. Second, this measure is independent of the mean, thus being useful in comparing variability between non-contiguous periods. For example, it would allow us to compare the volatility between two different years of a specific month. This approach therefore allows a more detailed and comparable analysis of the internal dynamics of the thermal signal, facilitating the study of recurrent patterns or local anomalies in climate variability. Before proceeding with the analysis of the results, it is appropriate to clarify the definition of volatility adopted in this study.

Definition 2.6.1. *Given a specific time interval* $[t_1, t_N]$*, the observed temperatures are* *denoted by T_i for $i = 1, \ldots, N$. The estimator of the temperature volatility σ over this interval is based on the quadratic variation of T_i and is defined as the squared root of:*

$$\hat{\sigma}^2_{[t_1,t_N]} = \frac{1}{N-1} \sum_{i=1}^{N-1} (T_{i+1} - T_i)^2. \tag{2.1}$$

A first step in the volatility analysis was to estimate σ^2_{year}, which represents the squared volatility calculated, according to Definition 2.6.1, on an annual basis for a specific city. This step was implemented with the aim of identifying any recurring or temporal trends year by year. The analysis of historical data relating to the period 2005–2024, reported in Figure 2.15, did not highlight significant trends that would justify a targeted in-depth analysis of the evolution of volatility. The annual values of σ^2 generally fluctuate over time, with no obvious systematic increases or decreases.

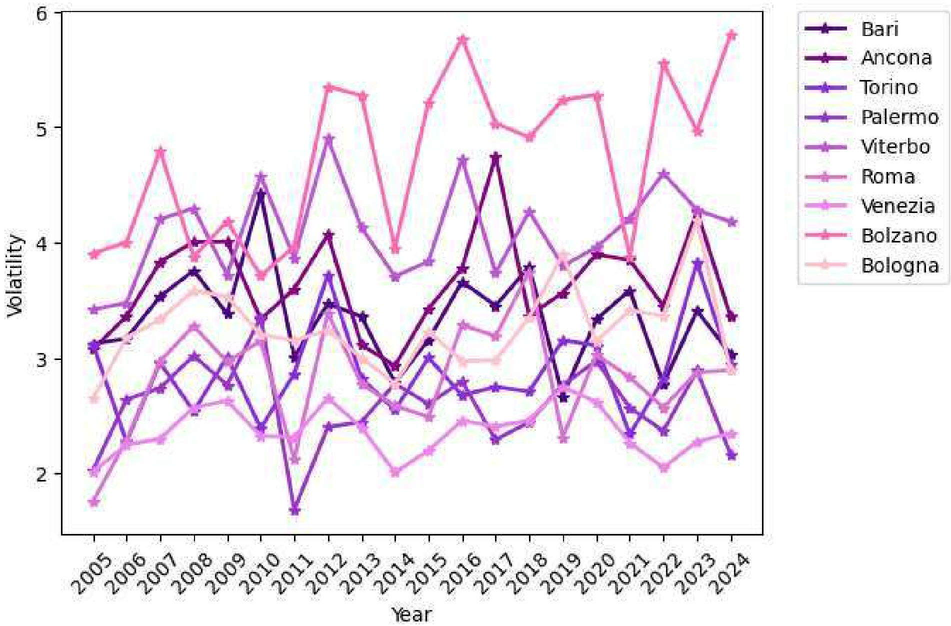
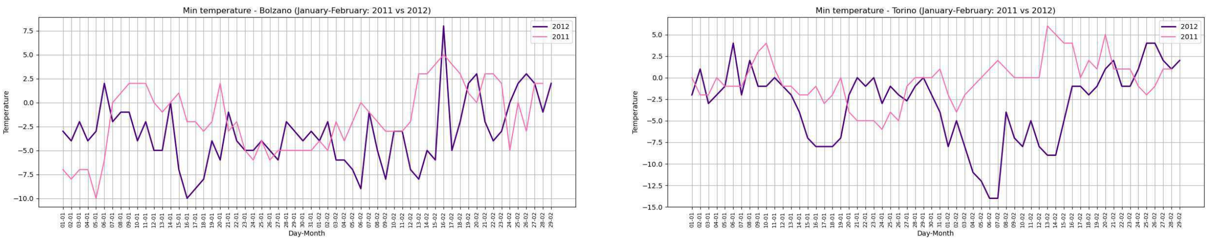


Figure 2.15: Annual variance estimated for each city.

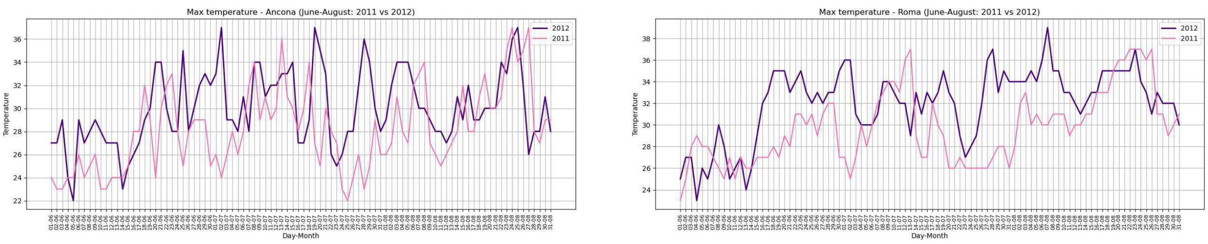
From Figure 2.15 some interesting observations can be made. First, the cities with the highest volatility appear to be those located near mountain systems, with Bolzano and Viterbo clearly standing out from the others. Secondly, volatility peaks are observed in correspondence with some specific years, which is common to most Italian cities. A significant example is 2012, a year in which volatility appears significantly higher than the averages of individual cities. A more in-depth analysis of the temperatures of that year reveals the presence of marked thermal anomalies: in particular, exceptionally low temperatures were recorded in the months of January and February, while the summer was characterized by values above the seasonal average. In Figure 2.16, several sudden peaks can be observed in 2012, indicative of significant temperature changes between consecutive days, which directly contribute to the increase in volatility, according to the adopted definition. Digging deeper, it is clear that 2012 was a meteorologically anomalous year: in February, cold waves from Russia hit much of the Eurasian continent. The same year also saw significant summer heat waves. This is clearly visible in the Figure 2.17 showing the maximum summer temperatures of 2012 and 2011 for the cities of Ancona and Rome. Again, in 2012, very pronounced daily variations can be noted, confirming the increase in observed volatility.



(a) Bolzano temperature in 2012 vs 2011

(b) Torino temperature in 2012 vs 2011

Figure 2.16: Torino and Bolzano temperature for January-February 2011-2012.



(a) Ancona temperature in 2012 vs 2011.

(b) Rome temperature in 2012 vs 2011.

Figure 2.17: Rome and Ancona temperature for June-August 2011-2012.

After the initial analysis of annual variance, illustrated in Figure 2.15, it was observed that this representation, although useful to provide a general picture of the trend of thermal variability, does not allow to clearly identify any recurring patterns or systematic anomalies over time. In the absence of marked or persistent trends, it was decided to adopt a more granular approach, exploring volatility according to a seasonal subdivision of the data. To this end, the dataset was grouped according to conventional meteorological seasons as already done in previous analyses, calculating for each season the variance of temperatures in the different cities. This type of analysis allowed to highlight significant differences between seasons, which would have remained hidden in an aggregated annual assessment. In particular, observing the Figure 2.18, a tendency towards greater volatility emerged during the intermediate seasons, i.e., spring and autumn. This behavior is consistent with what was expected from a meteorological point of view, as these seasons are characterized by a more marked transition between contrasting climatic conditions, which can favor sudden thermal changes. An interesting aspect that emerges from the figure concerns the different behaviors of temperature variance based on geographical position. In particular, it is possible to observe how coastal cities tend to show lower volatility than cities located inland or near mountainous reliefs. This behavior, already discussed previously and observed for the coastal cities of Rome, Palermo and Venice, is due to the mitigating influence of the seas that contribute to reducing the amplitude of daily temperature variations. On the contrary, Bolzano and Viterbo show significantly higher levels of volatility due to their position near mountainous reliefs. Bari and Ancona are located in an intermediate position between these two extremes. Despite being a coastal city, overlooking the Adriatic, they show greater volatility than other Tyrrhenian coastal cities, such as Rome. This apparent anomaly can be explained by considering the different conformation of the territory and the prevailing atmospheric circulation. The Adriatic, being a shallower and narrower basin than the Tyrrhenian, exerts a weaker mitigating effect. But this would not be enough to explain the different behavior of the two cities given the low volatility measured for the city of Venice. What makes the difference is the exposure of Bari and Ancona to cold flows from the Balkans, which can arrive directly without encountering significant orographic obstacles. On the contrary, cities located on the Tyrrhenian side benefit from the protection offered by the Apennines, which act as a natural barrier, limiting the entry of continental air masses from the north-east.

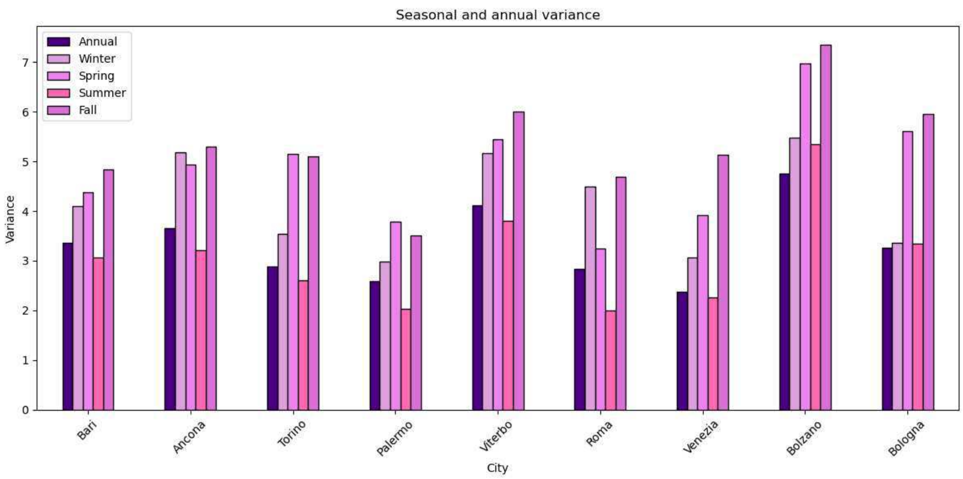


Figure 2.18: Annual and seasonal variance.

In order to further refine the variance analysis, it was decided to proceed with a monthly study. For each month, the variance relative to each year of the period considered was calculated, and the average value was subsequently determined by means of the arithmetic mean of the 20 values obtained. This choice, which will also be subsequently used in the definition of one of the models for the temperature trend, responds to the need to avoid that a seasonal aggregation, which includes months characterized by different climatic dynamics, can mask significant variations in volatility. An example is the case of April and May which, despite both belonging to the spring season, show significantly different behaviors, as can be seen from the Figure 2.19 showing the average temperatures of the city of Bologna in the two months for the years 2021 and 2024, with the month of May showing visibly lower volatility than the month of April. From the Figure 2.20

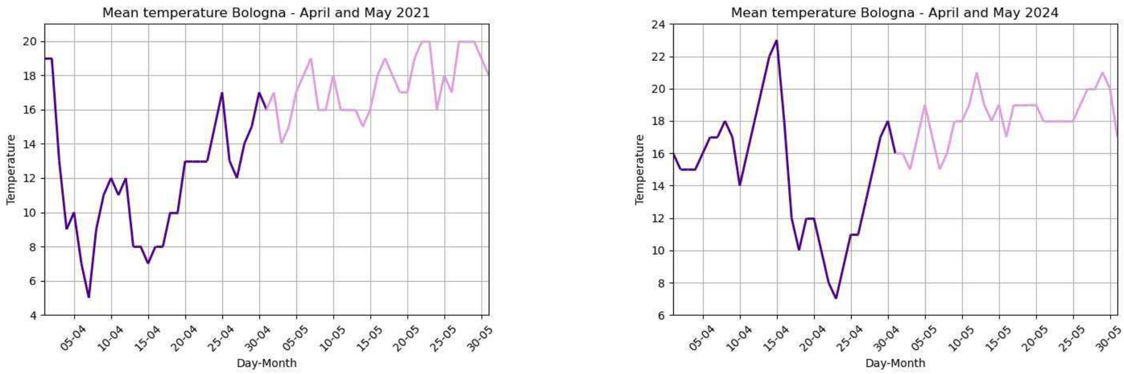


Figure 2.19: Bologna April and May temperature for 2021 and 2024.

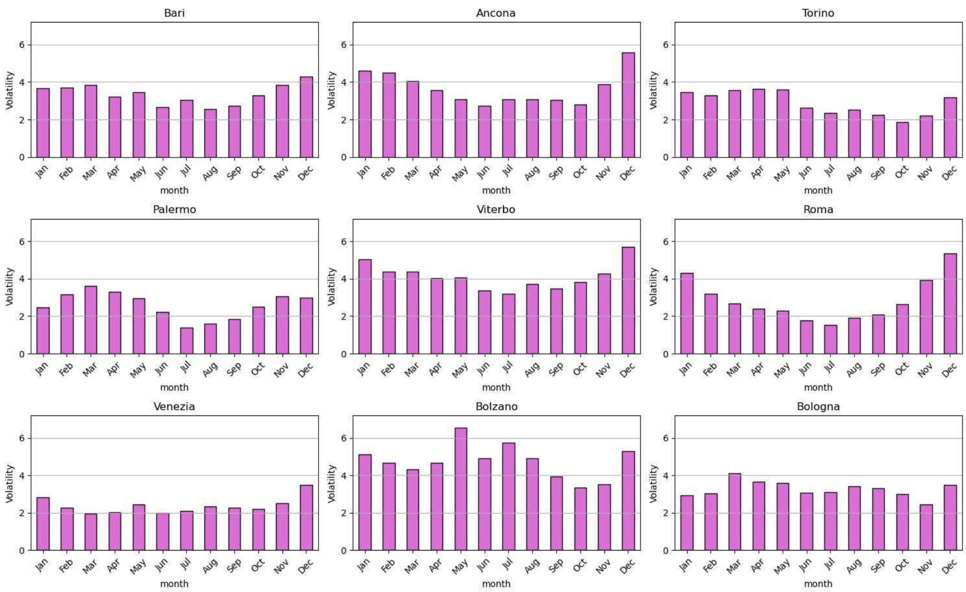


Figure 2.20: Monthly volatilities were computed using the Equation (2.1).

we can see how in relation the volatility shows a trend that can be well approximated with the seasonal values seen in the Figure 2.18 with minimum values concentrated in the summer months with a progressive increase in the autumn months culminating with a high volatility in the month of December. However, it is important to observe how the structure of the volatility trend depends on the climatic context of the reference city.

- Bari: presents a relatively contained and stable volatility for a good part of the year. This behavior is influenced by the presence of the Adriatic Sea. Compared to other coastal cities such as Venice and Palermo, the volatility of the Apulian city is slightly higher. The reason could be the cold currents coming from the Balkans.

- Ancona: presents a behavior similar to the city of Bari. What differentiates them slightly is the higher volatility of the city in the Marche due to the presence of the Apennines.

- Turin: presents a stable volatility during the seasons with a minimum in the summer season. An explanation for this behavior may be the presence of the barrier of the Alps that acts as a barrier to cold currents from Northern Europe.

- Palermo: presents with Venice a significantly lower volatility among the cities analyzed. This stability is consistent with the Mediterranean climate, where the mitigating influence of the sea reduces temperature excursions and seasonal atmospheric variations.

- Viterbo: presents a minimum volatility in the summer months and a sharp increase starting from September.

- Rome: presents a behavior similar to nearby Viterbo. Only a slight decrease in volatility is observed due to the proximity to the Tyrrhenian Sea.

- Venice: presents a relatively low volatility. The sub-lagoon climate, characterized by high humidity and gradual seasonal transitions, seems to confer greater uniformity to the volatility profile.

- Bolzano: appears to be an exception to what has been said in general about Italian cities. It shows anomalously high levels in summer. A cause can be found in the Föhn, a warm and dry wind that descends from the Alps towards the Bolzano valley causing a rapid warming of temperatures. In summer, this phenomenon can cause strong temperature fluctuations between two consecutive days, with a sudden increase in temperatures when the Föhn is active, followed by an equally rapid drop when the wind changes direction or stops. This thermal shock can explain the high volatility of temperatures observed on some summer days in Bolzano.

- Bologna: presents a moderate volatility with a regular trend. A seasonal pattern of volatility similar to that of Turin can be observed, although with less accentuated variations due to the greater proximity to the sea.

The analysis confirms that, within the Italian context, the local climate represents a determining factor in the modulation of monthly volatility. The cities of the South and of the coastal areas show a more contained and stable volatility, while the centers of the North and of the internal or mountainous areas show a more accentuated variability, particularly in the months of climatic transition.

In conclusion, given the absence of a clear trend in temperature fluctuations and considering that taking a seasonal mean value may not provide an adequate representation

	Annual	Winter	Spring	Summer	Fall
Bari	×	×	✓	×	×
Ancona	×	×	✓	×	×
Torino	×	×	✓	×	×
Palermo	×	×	×	×	×
Viterbo	×	×	✓	×	×
Roma	×	✓	×	×	×
Venezia	×	✓	×	×	×
Bolzano	×	×	×	×	×
Bologna	×	✓	×	×	×

Table 2.6: Normality for the seasonal and annual data.

	January	February	March	April	May	June	July	August	September	October	November	December
Bari	×	✓	×	×	✓	✓	×	✓	✓	✓	✓	×
Ancona	×	✓	✓	✓	×	✓	×	✓	✓	✓	✓	×
Torino	×	×	×	×	✓	✓	✓	✓	×	✓	×	×
Palermo	×	×	×	×	×	×	×	×	×	×	×	×
Viterbo	✓	×	×	✓	✓	×	×	✓	×	×	×	×
Roma	✓	×	×	✓	×	✓	×	✓	✓	×	×	✓
Venezia	✓	×	×	✓	✓	×	✓	×	×	×	×	✓
Bolzano	×	✓	×	×	×	×	×	×	×	✓	×	×
Bologna	✓	×	×	✓	✓	×	×	✓	×	×	×	×

Table 2.7: Normality for the monthly data.

of the variability, it was decided to adopt the mean of monthly volatilities as reference values. This choice allows obtaining a more accurate estimate of the specific thermal fluctuations for each month, taking into account the different dynamics that influence the temperature during the summer and improving the understanding of short-term thermal variations.

2.6.4 Temperature distribution

A final analysis carried out before proceeding with the modeling of the temperature, concerns the distribution of average temperatures. The analysis was initially conducted on all the values of the data set using a graphical approach as shown in Figure 2.21. As can be easily seen, the histograms of annual temperatures for the different cities analyzed show shapes that deviate significantly from the typical normal distribution curve. Asymmetries, the presence of multiple modes, heavier tails or flatter distributions compared to the normal one are observed. To confirm what was intuited from the graphs, it was decided to use statistical tests to confirm the non-normality of the data. For a more in-depth analysis, following what was already done with volatility, the distribution was explored more deeply by dividing the data at seasonal level. The results of the Jarque–Bera test for normality, reported in Table 2.6, show that in some seasons the average temperature actually follows a normal distribution, as could be intuited from Figure 2.22. The test is based on the sample skewness and kurtosis, making it particularly easy to interpret in terms of deviations from the symmetric, bell-shaped profile of the normal distribution. This suggests that a finer temporal subdivision allows us to identify conditions of normality that do not emerge at the annual level, even though there is still variability between locations and seasons. In light of these results, and consistently with what was done in the volatility analysis, we decided to proceed further with the subdivision up to the monthly detail. The monthly analysis of normality reported in the Table 2.7 showed even more favorable results: in several months, such as February, May, June, September, October and November, many cities showed a normal distribution of temperatures. However, even at the monthly level, Palermo stands out for the absence of normality in all months. This trend confirms that a greater temporal granularity favors compliance with the normality hypothesis and makes the monthly analysis particularly suitable for statistical applications that presuppose this condition.

	Annual		**Fall**		**Summer**		**Spring**		**Winter**	
City	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**	**Skewness**	**Kurtosis**
Bari	0.135050	-0.986416	0.028500	-0.560294	-0.198718	0.538800	0.070567	-0.186567	0.012149	0.487412
Ancona	0.039069	-1.036704	-0.050420	-0.465983	-0.261215	-0.031837	-0.017116	-0.264672	0.313208	0.384021
Torino	-0.041391	-1.087425	-0.090840	-0.739209	-0.227519	-0.140575	-0.098983	-0.111556	-0.003841	0.637342
Palermo	0.136781	-1.052322	0.052287	-0.416585	-0.222175	1.208916	0.366354	0.488374	0.326255	0.999181
Viterbo	0.137715	-0.971932	-0.162632	-0.551409	-0.406833	-0.017910	0.053545	-0.021710	-0.256730	0.010883
Roma	0.079193	-1.000218	-0.170943	-0.442079	-0.407894	-0.014715	0.135187	-0.150870	-0.082458	-0.005139
Venezia	-0.020201	-1.130344	-0.116265	-0.682571	-0.287662	-0.207112	-0.058407	-0.292065	-0.079861	-0.076326
Bolzano	-0.091807	-1.111186	-0.134182	-0.749594	-0.269082	-0.295108	-0.123331	-0.177334	0.239913	-0.036263
Bologna	-0.016249	-1.082250	-0.084201	-0.691870	-0.377356	-0.145341	-0.183487	-0.030079	-0.099423	0.074980

Table 2.8: Temperature Skewness and Kurtosis.

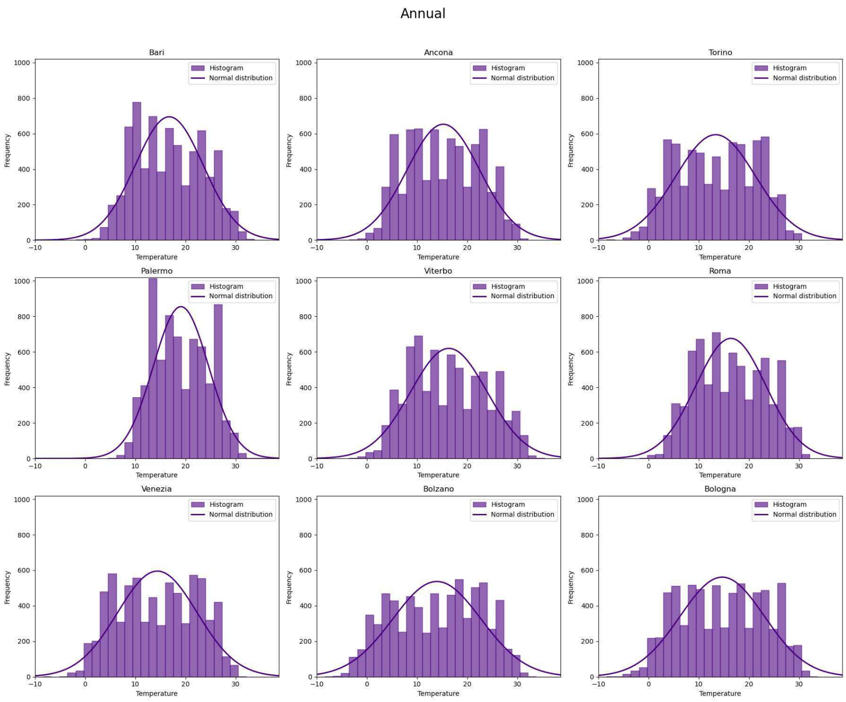


Figure 2.21: Distribution of annual temperature.

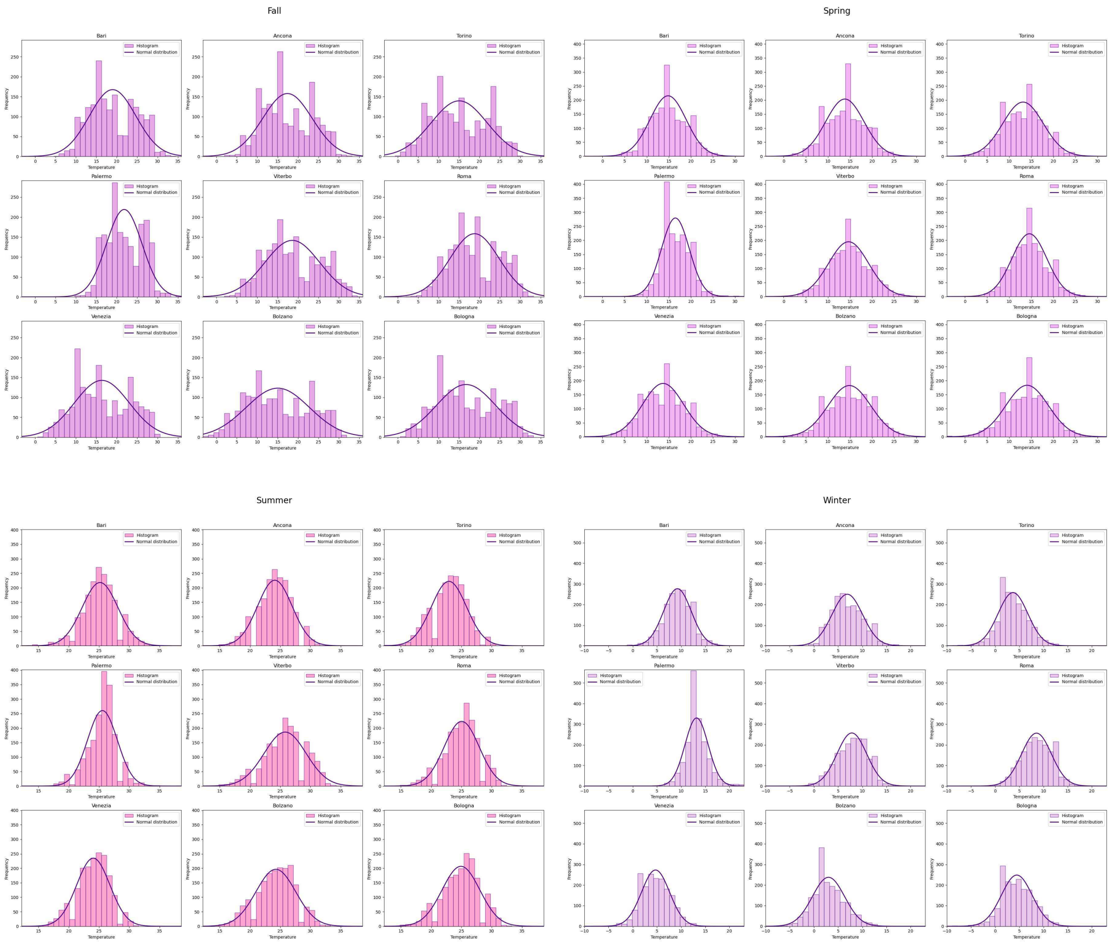


Figure 2.22: Distribution of seasonal temperature.

Chapter 3

Modeling temperature

"All models are wrong, but some are useful."
— George E. P. Box (1987)

In this chapter, the main objective is to build a model for the historical temperature series of different Italian cities. In a first phase, two distinct approaches will be introduced: on the one hand, an autoregressive moving average (ARMA) model, and on the other, a generalized Ornstein-Uhlenbeck process, in which the mean evolves over time. Subsequently, we will analyze how the length of the historical series used for the estimation of the parameters influences the predictive performances. In particular, the results obtained using the entire available series (20 years) will be compared with those deriving from the use of shorter time windows, equal to 10 and 5 years. The final objective will be to evaluate the effectiveness of the different models and different time lengths in producing forecasts for the year 2024, deliberately excluded from the calibration phase.

3.1 Deterministic component

Before proceeding with the stochastic modeling, it is essential to carry out an exploratory analysis of the available data, in order to identify any deterministic components, such as trends or seasonality, that must be correctly incorporated into the models. Temperatures, in particular, present a well-recognizable dynamic, strongly influenced by climatic seasonality and, to a lesser extent, by long-term variations. Figure 3.1 shows the time series of daily minimum, average and maximum temperatures observed in the city of Bari for the period 2005–2024. The 20-year interval allows us to appreciate the regularity with which seasonal trends are repeated, characterized by winter minimums and summer maximums. Zooming in on the year 2021 allows us to observe in greater detail the typical wave shape of temperatures during the calendar year. From a modeling point of view, this cyclicity can be effectively represented by periodic functions. In particular, a function of the

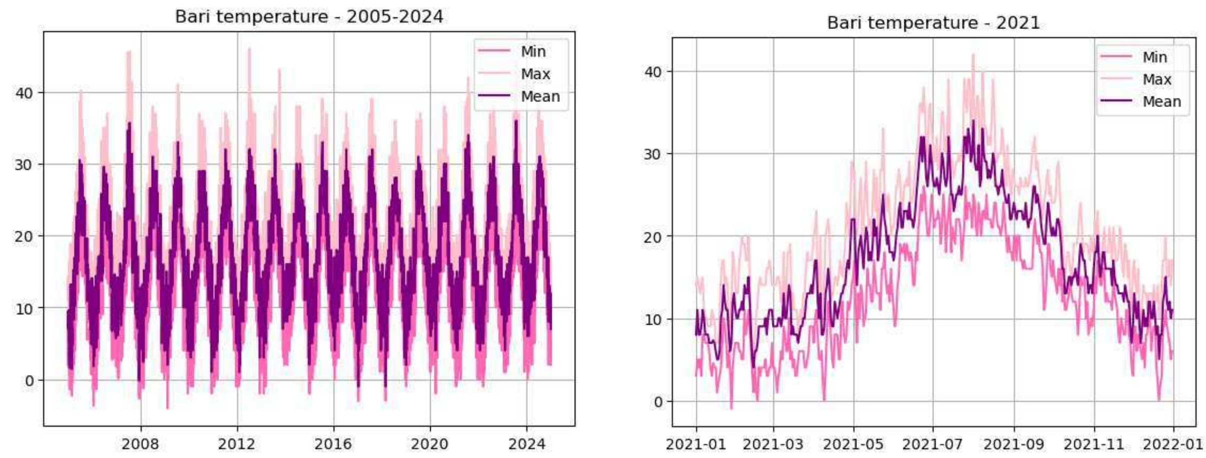


Figure 3.1: Average, minimum and maximum temperatures in Bari from 2005 to 2024 and a detailed view of the year 2021.

following type can be used:

$$f(t) = sin(\omega t + \phi),$$

where t represents the time expressed in days (with t=1 corresponding to January 1, 2005 and t=7304 to December 31, 2024). Considering that the observed periodicity is approximately one year, we set $\omega = \frac{2\pi}{365}$. The phase term ϕ allows us to take into account the time lag between the calendar and the actual temperature peaks, which do not necessarily coincide with the beginning or middle of the year. This correction is particularly relevant in the comparison between cities, since in coastal centers the mitigating influence of the sea tends to delay the occurrence of extreme values. The effect of the phase shift is illustrated in Figure 3.2, which compares mean temperatures in Bolzano and Palermo from 2022 to 2024, highlighting the lag in seasonal extremes observed. Figure 3.3, instead, provides a three-dimensional representation of the same data series, with the aim of simultaneously visualizing the evolution of temperatures within each year and between years. This type of visualization allows for a more immediate grasp of the periodic structure and relative stability of the seasonal cycle over time, also highlighting any anomalies or long-term trends.

From the descriptive analysis conducted in Chapter 2, the presence of an increasing trend in average daily temperatures over time clearly emerged, consistent with the scientific evidence relating to global warming and climate change observed in recent years. This trend represents a fundamental component to be taken into account in the modeling, as it reflects a systematic and non-random evolution of thermal behavior in the long term.

In addition to the seasonal component already analyzed, another essential element of the deterministic model is represented by a constant term, which allows to take into account the fact that the average annual temperature of a location does not oscillate around zero, but typically stabilizes at values around 15°C.

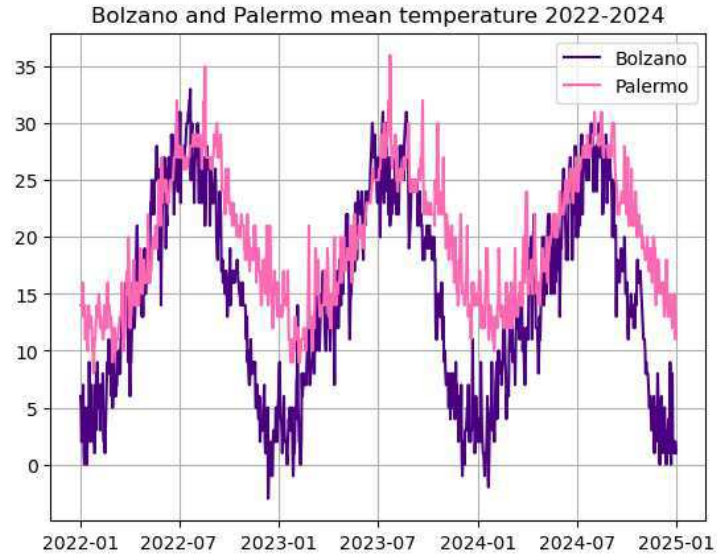


Figure 3.2: Daily mean temperatures in Bolzano and Palermo (2022–2024), showing the delayed seasonal peaks in the coastal city of Palermo.

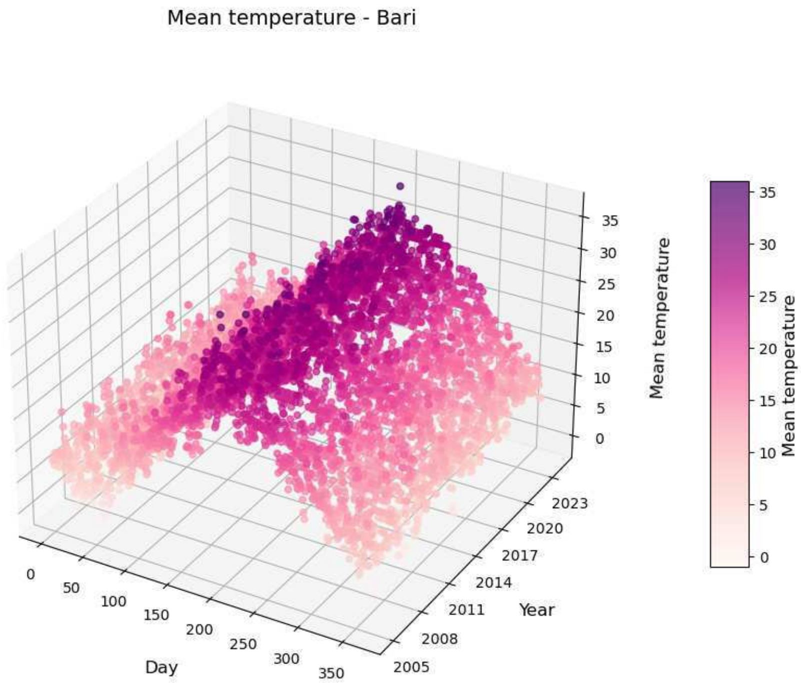


Figure 3.3: 3D plot for the average temperatures in Bari from 2005 to 2024.

Based on these considerations, we can therefore propose a deterministic model for the daily mean temperature at time t of the following form

$$T_t^m = A + B\,t + C\,\sin(\omega t + \phi) \tag{3.1}$$

where:

- The superscript m on T indicates that we are describing the average behavior over time,

- A represents the mean level,
- B is the coefficient of the linear trend,
- C is the amplitude of the seasonal oscillation,
- $\omega = \frac{2\pi}{365}$ represents the annual frequency,
- ϕ is a phase parameter that allows to model the phase shift between the maximum/minimum of the sine function and those observed in the real data.

The parameters A, B, C and ϕ will be estimated in order to obtain the best possible adherence to the historical data.

3.1.1 Parameter estimation

From now on, all parameter estimates for the models will be performed using data from 2005 to 2023, thus leaving the year 2024 exclusively for backtesting the models. To obtain the numerical values of the constants in the Equation (3.1), we rely on the least squares methodology, as suggested by [1]. The least squares method is a statistical technique commonly used to estimate the parameters of a mathematical model so that the model best fits the observed data. The goal of this method is to find the parameters that minimize the sum of the squares of the differences between the actual data and those predicted by the model. To estimate the model parameters we will rely on the equation

$$Y_t = a_1 + a_2 t + a_3 sin(\omega t) + a_4 cos(\omega t). \tag{3.2}$$

To determine the optimal values of these parameters, collected in the vector $v = (a_1, a_2, a_3, a_4)$, it is necessary to solve an optimization problem that minimizes the discrepancy between the observed data and those provided by the model. In particular, we want to minimize the quadratic error between the real average temperatures, indicated by X, and the values estimated by the model, indicated by Y. This leads to the following problem:

$$\begin{aligned}\min_v \|Y - X\|^2 &= \min_v \sum_{i=1}^{n} (Y_i - X_i)^2 \\ &= \min_v \sum_{t=1}^{n} \left((a_1 + a_2 t + a_3 \sin(\omega t) + a_4 \cos(\omega t)) - X_t\right)^2\end{aligned}$$

where the norm indicates the sum of the squares of the differences between the observed and estimated values. Once the parameter vector $v = (a_1, a_2, a_3, a_4)$ has been estimated, we can retrieve the constants in equation (3.1). Specifically, the constants A and B correspond directly to a_1 and a_2, respectively.

To determine the values of C and ϕ, a further transformation is required. Starting from equation (3.1), we can rewrite the sine term using the angle sum identity:

$$C\sin(\omega t+\phi)=C\left[\sin(\omega t)\cos(\phi)+\cos(\omega t)\sin(\phi)\right].$$

Comparing with equation (3.2), we obtain:

$$a_3=C\cos\phi \quad \text{and} \quad a_4=C\sin\phi.$$

To compute C from the estimated values of a_3 and a_4, we apply the Pythagorean identity:

$$a_3^2+a_4^2=C^2\cos^2(\phi)+C^2\sin^2(\phi)=C^2,$$

thus:

$$C=\sqrt{a_3^2+a_4^2}.$$

To determine the phase parameter ϕ, we take the ratio:

$$\frac{a_4}{a_3}=\frac{C\sin(\phi)}{C\cos(\phi)}=\tan(\phi),$$

which gives:

$$\phi=\arctan\left(\frac{a_4}{a_3}\right)-\pi.$$

The subtraction of π is a modeling choice intended to place the phase angle ϕ in the desired quadrant. This is particularly useful when modeling seasonal effects, as the timing of temperature peaks and troughs may vary depending on geographic factors, such as the moderating influence of the sea in coastal cities. In summary, once the vector values v have been obtained, the model constants can be derived using the following formulas:

- $A=a_1$;
- $B=a_2$;
- $C=\sqrt{a_3^2+a_4^2}$;
- $\phi=\arctan\left(\frac{a_4}{a_3}\right)-\pi$.

The estimated values for vector v are reported in Table 3.1 and the corresponding parameters in Table 3.2. In Table 3.1 is also reported the coefficient of determination, R^2. This statistic quantifies the proportion of the variance in the observed values X that is accounted for by the fitted model Y. An R^2 value approaching 1 signifies a model that closely fits the data, indicating a high explanatory power.

	Bologna	**Bari**	**Bolzano**	**Viterbo**	**Roma**	**Palermo**	**Ancona**	**Venezia**	**Torino**
const	13.903	15.969	12.956	15.780	15.472	18.749	14.528	13.809	12.351
t	0.000214	0.000212	0.000282	0.000151	0.000251	0.000088	0.000188	0.000158	0.000287
$\sin(\omega t)$	-3.722	-4.002	-2.484	-4.051	-4.078	-4.397	-3.824	-3.616	-3.217
$\cos(\omega t)$	-10.514	-7.870	-11.372	-9.041	-8.146	-5.747	-8.639	-9.944	-10.023
R^2	0.880	0.851	0.879	0.848	0.861	0.862	0.859	0.893	0.880

Table 3.1: Estimated parameters for the temperature model in each city.

	Bologna	**Bari**	**Bolzano**	**Viterbo**	**Roma**	**Palermo**	**Ancona**	**Venezia**	**Torino**
A	13.903	15.969	12.956	15.780	15.472	18.749	14.528	13.809	12.351
B	0.000214	0.000212	0.000282	0.000151	0.000251	0.000088	0.000188	0.000158	0.000287
C	11.154	8.829	11.640	9.907	9.110	7.236	9.447	10.582	10.526
ϕ	-1.911	-2.041	-1.786	-1.992	-2.035	-2.224	-1.987	-1.920	-1.881

Table 3.2: Estimated values of the constants A, B, C, and ϕ for each city.

In Figure 3.4 it is possible to see the behavior of the deterministic model with respect to the observed data. Before continuing, it was useful to observe the behavior of the

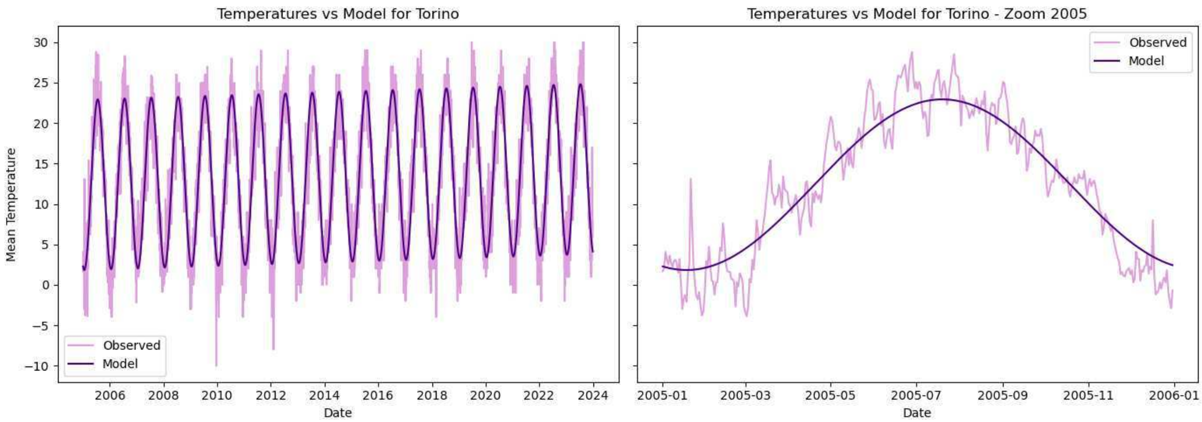


Figure 3.4: Deterministic model vs observed data for Turin.

series once the deterministic contribution was removed from the observed series, thus obtaining a series of residuals that represent the stochastic component not explained by the seasonal model. In Figure 3.5 we can observe the irregular fluctuations around zero, consistent with the hypothesis of a centered stochastic process. For a more detailed visualization of the dynamics on a reduced time scale, a zoom-in relating to the year 2005 has also been included, reported in the same figure, which highlights the alternation of short positive and negative deviations, typical of a mean-reverting behavior. To further assess the nature of the residuals, several diagnostic plots have been included in Figure 3.6. The autocorrelation function plot shows a gradual decay, suggesting the presence of correlations over time. The partial autocorrelation function highlights two significant lags, indicating potential dependencies at these specific time points. The QQ-plot reveals deviations from normality, particularly at the tails, which suggests that the residuals may not fully follow a Gaussian distribution.

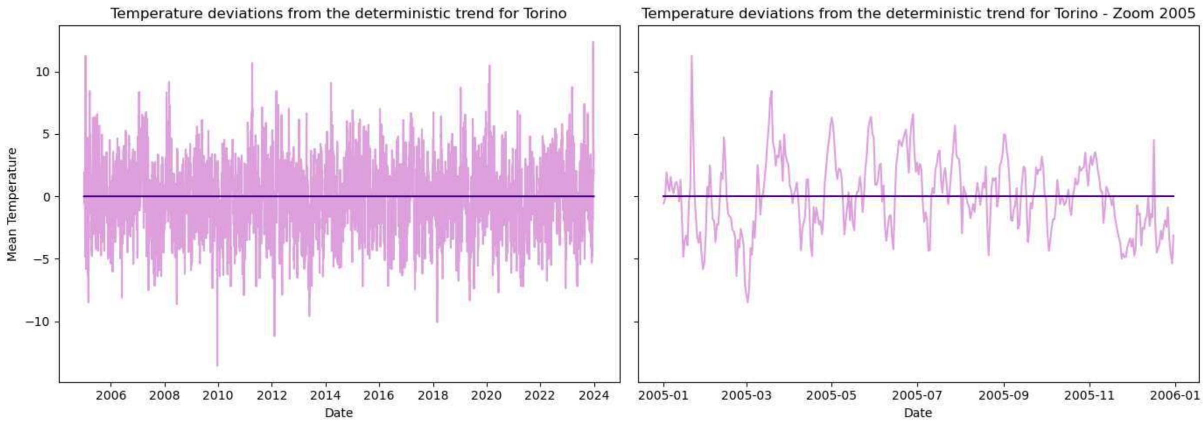


Figure 3.5: Stochastic component for Turin.

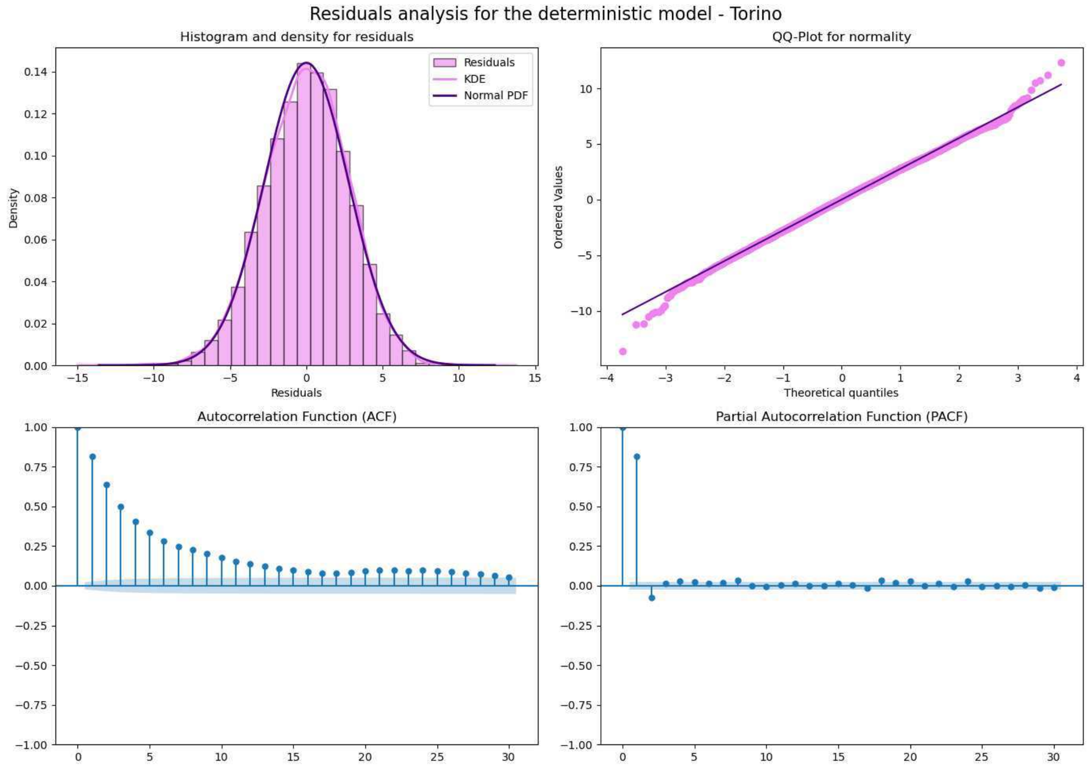


Figure 3.6: Residual analysis from deterministic model for Turin.

3.1.2 Impact of historical window length on model parameters

An important aspect in time series modeling is the choice of the amount of historical data to be used for model training. It must be considered that the size of the data window can influence the estimation of the parameters and consequently the predictive capabilities of the model. In this paragraph we want to propose a comparative analysis between three different scenarios. We will compare the models defined using historical series of length 5, 10 and 19 years of data respectively. The comparison is important to understand how to best balance two opposing needs. On the one hand, a longer time series guarantees greater statistical stability and a more robust representation of long-term dynamics, on the other, more recent periods could better capture any emerging structural changes. For each of the time horizons, the same process is applied entirely to evaluate their predictive capacity on the year 2024. The objective is to identify the time window that offers the best predictive model.

Parameter	Model	Bologna	Bari	Bolzano	Viterbo	Roma	Palermo	Ancona	Venezia	Torino
	Value	13.90	15.97	12.96	15.78	15.47	18.75	14.53	13.81	12.35
A	Δ 10 years [%]	6.10	6.40	12.03	2.60	5.02	1.64	5.69	6.84	8.47
	Δ 5 years [%]	6.47	5.72	7.69	4.94	6.44	0.29	5.33	4.59	10.59
	Value	0.000214	0.000212	0.000282	0.000151	0.000251	0.000088	0.000188	0.000158	0.000287
B	Δ 10 years [%]	-29.20	-85.71	-109.56	43.49	11.80	-1.71	-46.11	-123.49	-8.05
	Δ 5 years [%]	94.94	14.62	72.99	55.80	130.00	509.66	70.80	50.80	72.55
	Value	11.15	8.83	11.64	9.91	9.11	7.24	9.45	10.58	10.53
C	Δ 10 years [%]	-3.05	-0.53	0.06	-2.45	-2.37	0.43	-1.75	-1.93	-1.86
	Δ 5 years [%]	-3.01	1.63	0.74	-1.59	0.99	4.04	-0.86	-1.02	-1.24
	Value	-1.91	-2.04	-1.79	-1.99	-2.03	-2.22	-1.99	-1.92	-1.88
ϕ	Δ 10 years [%]	-0.32	-0.00	-0.65	-0.16	-0.07	0.07	-0.03	-0.49	-0.41
	Δ 5 years [%]	1.02	2.39	-0.62	1.06	0.84	0.30	1.69	0.94	-0.00

Table 3.3: Estimated parameters of the deterministic model with percentage changes compared to the reference model (19 years).

By looking at the Table 3.3 it is possible to understand the changes in the parameters of the deterministic model in percentage with respect to the value taken as a reference represented by the model discussed up to this point that takes into account all 19 years of available data. Evaluating the behavior of these parameters on three time horizons allows us to evaluate the structural stability of the model and understand if there have been disruptive evolution in the last years. The results show that the parameter A, associated with the constant component of the model, tends to increase in most cities in the 5 and 10-year models, suggesting an increase in the average temperatures in recent years. The variations are on average between 2% and 10%, with some higher peaks in the 5-year model. The B parameter, which captures the linear trend, shows a more unstable behavior: while in the 10-year model marked reductions or even inversions of sign are observed in many cities (indicating a weakening of the growth trend), the 5-year model instead shows very high increases, sometimes exceeding 100% or even 500% (as in the case of Palermo), indicating a potential strengthening of the recent trend. However, these increases must be interpreted with caution, given the greater sensitivity of the with such small numbers. The seasonal cyclical component (C) instead appears more stable: the variations with respect to the 19-year model are generally modest, oscillating around $\pm 3\%$, suggesting that the amplitude of the seasonal cycle has not changed significantly. The ϕ parameter, which regulates the phase shift of the seasonal cycle, also shows limited variations, indicating that the temporal position of the seasonal peaks has remained almost unchanged. Overall, this analysis highlights a good stability of the deterministic model, with significant changes mainly in the constant component and in the linear trend, which could reflect an evolution of the climate regime, while the seasonal structure seems to be preserved over time.

3.2 Model 1: Modified Ornstein-Uhlenbeck process

In the previous section, a deterministic model for temperature trends was introduced. Although this describes the temperature trend in time in a discrete way by incorporating seasonality, it does not capture a fundamental characteristic of the data. In fact, the described model does not take into account the presence of random fluctuations around the average trend represented by the deterministic model. From what we have also observed in Figure 3.4, the temperature does not follow a regular path but is subject to multiple random factors. These variations that are not precisely predictable and often sudden in nature cause the value of the daily average temperature to deviate, even significantly, from the expected average value for a certain day of the year. It follows that a purely deterministic model that does not allow fluctuations is insufficient when you want to describe the temperature dynamics more faithfully. To fill this gap, a stochastic model is introduced that is able to explicitly represent the uncertainty and the random component of the phenomenon. To describe the random behavior of temperatures, among the different approaches, it has been considered to use the Ornstein-Uhlenbeck (OU) process, a continuous-time stochastic process widely used to model random variables that show a tendency to return to a mean value. The OU process is one of the simplest solutions among the mean-reverting processes, it was originally introduced by Leonard Ornstein and George Uhlenbeck in the context of statistical physics to describe the motion of particles subjected to friction forces and thermal noise. In the case of our time-series, the OU process allows the representation of the temperature evolution as the result of the sum of two components: a tendency towards a mean value and a random component that introduces noise into the system. In the literature, it is common to distinguish between a deterministic and a stochastic component in temperature models. The deterministic component has a structure similar to the one illustrated in Section 3.1. As for the stochastic part, there are several approaches; one of them is the one presented in this section and described in [1] where Alaton proposes a model for temperatures using a stochastic process similar to the Ornstein-Uhlenbeck but with a slight modification. The mean towards which the process converges is not a fixed value but a deterministic function that describes the average anadmanence of temperatures. In this work a similar model is proposed for Italian temperatures

3.2.1 Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck, well describend in [14] process is a continuous-time stochastic process that serves as a classical model for mean-reverting dynamics. Introduced with the aim of describing physical phenomena, such as Brownian motion with friction, it is now widely used in the financial field. The ability to describe phenomena whose behavior is not purely random, but has a long-term structure has made it very popular. One of its most significant applications is in the modeling of short-term interest rates, as in the well-known Vasicek model, in which the rate evolves according to an OU process. This approach allows us to overcome the limitations of geometric Brownian motion, providing a more realistic dynamics, especially with regard to the possibility of negative values and the stationarity of the distribution.

Definition 3.2.1. *The OU process is defined as the solution to the following linear stochastic differential equation (SDE):*

$$dX_t = a(\mu - X_t)\,dt + \sigma\,dW_t, \quad X_0 = x \tag{3.3}$$

where:

- $\mu \in \mathbb{R}$ *is the long-term mean,*
- $a > 0$ *is the rate of mean reversion,*
- $\sigma > 0$ *is the volatility coefficient,*
- $\{W_t\}_{t\geq 0}$ *denotes a standard Brownian motion,*
- x *is a constant.*

The equation describes a process that experiences a pull toward the mean value μ with strength proportional to the deviation $\mu - X_t$, while being subjected to continuous stochastic perturbations introduced by the Brownian motion.

Theorem 3.2.1. *The explicit solution of the OU process described in (3.3) has the following form*

$$X_t = \mu + e^{-a(t-s)}(X_s - \mu) + \sigma \int_s^t e^{-a(t-u)} dW_u. \tag{3.4}$$

Proof. To solve the SDE we can simplify the equation (3.3) by introducing a change of variable. Let $Y_t = X_t - \mu$, then $X_t = Y_t + \mu$ and $dX_t = dY_t$. Substituting into the SDE:

$$\begin{aligned} dY_t &= a(\mu - (Y_t + \mu))dt + \sigma dW_t \\ &= a(-Y_t)dt + \sigma dW_t = -aY_t dt + \sigma dW_t \end{aligned}$$

Consider, now, the equation for Y and multiply both sides by the factor e^{at}:

$$e^{at} dY_t + a e^{at} Y_t dt = \sigma e^{at} dW_t$$

The left side is the differential of $e^{at}Y_t$:

$$d(e^{at}Y_t) = \sigma e^{at} dW_t$$

Integrating both sides from s to t:

$$\int_s^t d(e^{au}Y_u) = \int_s^t \sigma e^{au} dW_u$$

$$e^{at}Y_t - e^{as}Y_s = \int_s^t \sigma e^{au} dW_u$$

Isolating Y_t:

$$Y_t = e^{-a(t-s)}Y_s + e^{-at}\int_s^t \sigma e^{au} dW_u$$

Substituting $Y_t = X_t - \mu$ and $Y_s = X_s - \mu$:

$$X_t - \mu = e^{-a(t-s)}(X_s - \mu) + e^{-at}\int_s^t \sigma e^{au} dW_u$$

$$X_t = \mu + e^{-a(t-s)}(X_s - \mu) + e^{-at}\int_s^t \sigma e^{au} dW_u$$

Rearranging the terms:

$$\begin{aligned} X_t &= X_s e^{-a(t-s)} - \mu e^{-a(t-s)} + \mu + e^{-at}\int_s^t \sigma e^{au} dW_u \\ &= X_s e^{-a(t-s)} + \mu(1 - e^{-a(t-s)}) + \sigma \int_s^t e^{-a(t-u)} dW_u \end{aligned}$$

which is the solution of the Ornstein-Uhlenbeck process.

□

Before continuing, it is necessary to formalize the filtration on which we will work. For an Ornstein-Uhlenbeck process, denoted by $X = (X_t)_{t\geq 0}$, the most immediate filtration that can be defined is the natural one that is constructed directly from the observations of the process itself. Given the existence of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the OU process is defined, the natural filtration $(\mathcal{F}_t)_{t\geq 0}$ is a family of σ-algebras in which, for each instant $t \geq 0$, $\mathcal{F}_t$ is defined as the smallest σ-algebra that makes all the random variables X_s measurable for $0 \leq s \leq t$. It can be defined as follows:

$$\mathcal{F}_t = \sigma(\{X_s : 0 \leq s \leq t\}). \tag{3.5}$$

The $\mathcal{F}_t$ filter as such captures the entire history of the OU process up to time t, ensuring that every observable event up to that time is contained in the information flow.

The intrinsic non-decreasing property ($\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s < t$), ensures that information progressively accumulates without ever decreasing. Crucially, this filtering makes the process X_t adaptive, meaning that the value of X_t is always measurable with respect to $\mathcal{F}_t$.

Proposition 3.2.1. *The Ornstein-Uhlenbeck process $\{X_t\}_{t\geq 0}$, defined as the solution to the stochastic differential equation*

$$dX_t = -a(X_t - \mu)\, dt + \sigma\, dW_t,$$

with parameters $a > 0$, $\mu \in \mathbb{R}$, $\sigma > 0$, and W_t a standard Brownian motion, is a Markov process.

Proof. The solution to the SDE with initial condition $X_0 = x$ can be expressed explicitly as

$$X_t = \mu + (X_0 - \mu)e^{-at} + \sigma \int_0^t e^{-a(t-s)}\, dW_s.$$

For any $0 \leq s < t$, this can be rewritten as

$$X_t = \mu + (X_s - \mu)e^{-a(t-s)} + \sigma \int_s^t e^{-a(t-u)}\, dW_u.$$

The integral $\int_s^t e^{-a(t-u)}\, dW_u$ depends only on the increments of the Brownian motion after time s, and thus is independent of the sigma-algebra $\mathcal{F}_s$ generated by the process up to time s. Therefore, the conditional distribution of X_t given the entire past $\mathcal{F}_s$ depends only on the present state X_s:

$$\mathbb{P}(X_t \in A \mid \mathcal{F}_s) = \mathbb{P}(X_t \in A \mid X_s),$$

for every measurable set A. This verifies the Markov property for the Ornstein-Uhlenbeck process. □

Observation 3.2.1. *Since the process is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, and X_s is measurable with respect to $\mathcal{F}_s$, the sigma-algebra generated by X_s, denoted $\sigma(X_s)$, is a sub-sigma-algebra of $\mathcal{F}_s$. Thus, conditioning on $\mathcal{F}_s$ can be replaced by conditioning on $\sigma(X_s)$ because the future evolution depends only on the current state X_s. This replacement simplifies the Markov property expression and highlights that the process is memoryless beyond its current state.*

Remark. Throughout this work, we write expressions such as $\mathbb{E}[f(X_t) \mid X_s]$ to denote conditional expectations given the information contained in the value of the process at time s. Formally, this corresponds to conditioning with respect to the σ-algebra generated by X_s, that is, $\sigma(X_s)$. We adopt this simplified notation to avoid confusion with the parameter σ, which is already used to denote the volatility of the process.

Proposition 3.2.2. *We can compute explictly the mean, the variance and the covariance of the process:*

- *Mean:*

$$
\begin{aligned}
E[X_t|X_s] &= E\left[X_s e^{-a(t-s)} + \mu(1-e^{-a(t-s)}) + \sigma \int_s^t e^{-a(t-u)} dW_u \Big| X_s\right] \\
&= X_s e^{-a(t-s)} + \mu(1-e^{-a(t-s)}) + \sigma E\left[\int_s^t e^{-a(t-u)} dW_u \Big| X_s\right] \\
&= X_s e^{-a(t-s)} + \mu(1-e^{-a(t-s)}) + \sigma \cdot 0 \\
&= X_s e^{-a(t-s)} + \mu(1-e^{-a(t-s)})
\end{aligned}
$$

Asymptotic behavior:

$$\lim_{t\to\infty} E[X_t|X_s] = \mu$$

- *Variance:*

$$Var(X_t|X_s) = E[(X_t - E[X_t|X_s])^2|X_s], \qquad X_t - E[X_t|X_s] = \sigma \int_s^t e^{-a(t-u)} dW_u$$

$$
\begin{aligned}
Var(X_t|X_s) &= E\left[\left(\sigma \int_s^t e^{-a(t-u)} dW_u\right)^2 \Big| X_s\right] = \sigma^2 E\left[\left(\int_s^t e^{-a(t-u)} dW_u\right)^2 \Big| X_s\right] \\
&= \sigma^2 \int_s^t \left(e^{-a(t-u)}\right)^2 du \\
&= \sigma^2 \int_s^t e^{-2a(t-u)} du
\end{aligned}
$$

Replacing $v = -2a(t-u)$*,* $du = \frac{dv}{2a}$*:*

$$
\begin{aligned}
Var(X_t|X_s) &= \sigma^2 \int_{-2a(t-s)}^0 e^v \frac{dv}{2\theta} = \frac{\sigma^2}{2a}[e^v]_{-2a(t-s)}^0 \\
&= \frac{\sigma^2}{2a}(1 - e^{-2a(t-s)})
\end{aligned}
$$

Asymptotic behavior:

$$\lim_{t\to\infty} Var(X_t|X_s) = \frac{\sigma^2}{2a}$$

- *Covariance: for* $0 \leq s \leq t \leq u$,

$$\begin{aligned}
Cov\left[(X_t, X_u)|X_s\right] &= \mathbb{E}\left[(X_t - \mathbb{E}[X_t])(X_u - \mathbb{E}[X_u])|X_s\right] \\
&= \mathbb{E}\left[\left(\int_0^t \sigma e^{a(v-t)} dW_v\right)\left(\int_0^u \sigma e^{a(w-u)} dW_w\right)|X_s\right] \\
&= \sigma^2 e^{-a(t+u)} \mathbb{E}\left[\left(\int_0^t e^{av} dW_v\right)\left(\int_0^u e^{aw} dW_w\right)\right] \\
&= \sigma^2 e^{-a(t+u)} \mathbb{E}\left[\int_s^t e^{2av} dv|X_s\right] = \sigma^2 e^{-a(t+u)}\left(e^{2at} - e^{2as}\right),
\end{aligned}$$

where we used that the part of the integral from t *to* u *does not contribute to the conditional covariance because it involves Brownian increments after time* t*, which are independent of everything up to time* t*. Similarly, the part from* 0 *to* s *does not matter because it is fully measurable with respect to* X_s*, and thus behaves like a constant when conditioning on* X_s*. Therefore, only the stochastic components from* s *to* t *affect the conditional covariance.*

Proposition 3.2.3. *Let* X_t *be the solution to the Ornstein–Uhlenbeck stochastic differential equation*

$$dX_t = -a(X_t - \mu)\, dt + \sigma\, dW_t, \qquad \text{with initial condition } X_s = x \in \mathbb{R} \text{ deterministic.}$$

Then for every $t \geq s$*, the random variable* X_t *is normally distributed with mean* $X_s e^{-a(t-s)} + \mu(1 - e^{-a(t-s)})$ *and variance* $\frac{\sigma^2}{2a}(1 - e^{-2a(t-s)})$*.*

Proof. The Ornstein–Uhlenbeck process has the explicit solution:

$$X_t = \mu + e^{-a(t-s)}(X_s - \mu) + \sigma \int_s^t e^{-a(t-u)} dW_u.$$

Here, the first two terms are deterministic for every $t \geq s$, while the stochastic integral

$$\int_s^t e^{-a(t-u)}\, dW_u$$

is a Itô integral of a deterministic function with respect to Brownian motion. Such an integral is normally distributed with mean zero and finite variance given by

$$Var\left(\int_s^t e^{-a(t-u)}\, dW_u\right) = \int_s^t e^{-2a(t-u)}\, du = \frac{1 - e^{-2a(t-s)}}{2a}.$$

Therefore, X_t is the sum of deterministic terms and a normally distributed term, hence it is itself normally distributed. □

3.2.2 Modified Ornstein-Uhlenbeck process

The initial approach was to model the temperature dynamics using a standard Ornstein–Uhlenbeck process. However, it became clear that the mean of the temperature process described previously is not constant over time, making the standard formulation given in Definition 3.2.1 unsuitable. To address this mismatch, an additional term is introduced into equation (3.3), leading to the modified formulation presented below.

$$dT_t = dT_t^m + a(T_t^m - T_t)dt + \sigma dW_t \tag{3.6}$$

with starting point $T_s = x$ and $t > s$. The process defined in Equation (3.6) can be seen as a generalization of the standard Ornstein–Uhlenbeck process described in Definition 3.2.1. Specifically, if the deterministic function T_t^m is constant over time, say $T_t^m \equiv \mu$, then the differential dT_t^m vanishes and the equation reduces to:

$$dT_t = a(\mu - T_t)\,dt + \sigma\,dW_t,$$

which corresponds exactly to the classical OU dynamics. In this modified formulation, the mean T_t^m becomes a deterministic, time-varying target level towards which the stochastic process T_t reverts. The additional drift term dT_t^m accounts for the evolution of the deterministic component, such as seasonal cycles or long-term trends. This modification allows the model to preserve the mean-reversion structure of the OU process while adapting to non-stationary behavior observed in real temperature series.

Proposition 3.2.4. *The solution of the process described in Equation (3.6) is the following*

$$T_t = (x - T_s^m)e^{-a(t-s)} + T_t^m + \sigma \int_s^t e^{-a(t-u)} dW_u. \tag{3.7}$$

Proof. To find the solution to the Equation (3.6) it is helpful to introduce the change of variable

$$X_t = T_t - T_t^m \quad \Rightarrow \quad dX_t = dT_t - dT_t^m.$$

Substituting in Equation 3.6 we obtain

$$dX_t = dT_t - dT_t^m = a(T_t^m - T_t)dt + \sigma dW_t = -aX_t dt + \sigma dW_t. \tag{3.8}$$

We can observe that the Equation (3.8) corresponds to the OU process described in Equation (3.3) with the costant mean $\mu = 0$. So we can apply the Theorem 3.2.1 and obtain the solution for the process X_t:

$$X_t = e^{-a(t-s)} X_s + \sigma \int_s^t e^{-a(t-u)} dW_u.$$

Substituting $X_t = T_t - T_t^m$ and $T_s = x$ we obtain the solution for the modified OU process

$$T_t - T_t^m = e^{-a(t-s)}(T_s - T_s^m) + \sigma \int_s^t e^{-a(t-u)} dW_u,$$

$$T_t = T_t^m + e^{-a(t-s)}(x - T_s^m) + \sigma \int_s^t e^{-a(t-u)} dW_u.$$

$\square$

To highlight the mean reversion property of the modified Ornstein-Uhlenbeck process considered, the Figure 3.7 reports the plot of the mean of n simulations of the process. Each simulation represents a possible path of the stochastic process, influenced by Brownian noise. The simulations were performed using the temperature model calibrated for the city of Venice, solving the SDE numerically. However, by calculating the mean on an increasingly larger number of trajectories, the random contribution tends to compensate, and the empirical mean converges towards the model. This behavior visually confirms the tendency of the process to return to its mean over time, even when the latter varies according to a trend and a periodic component.

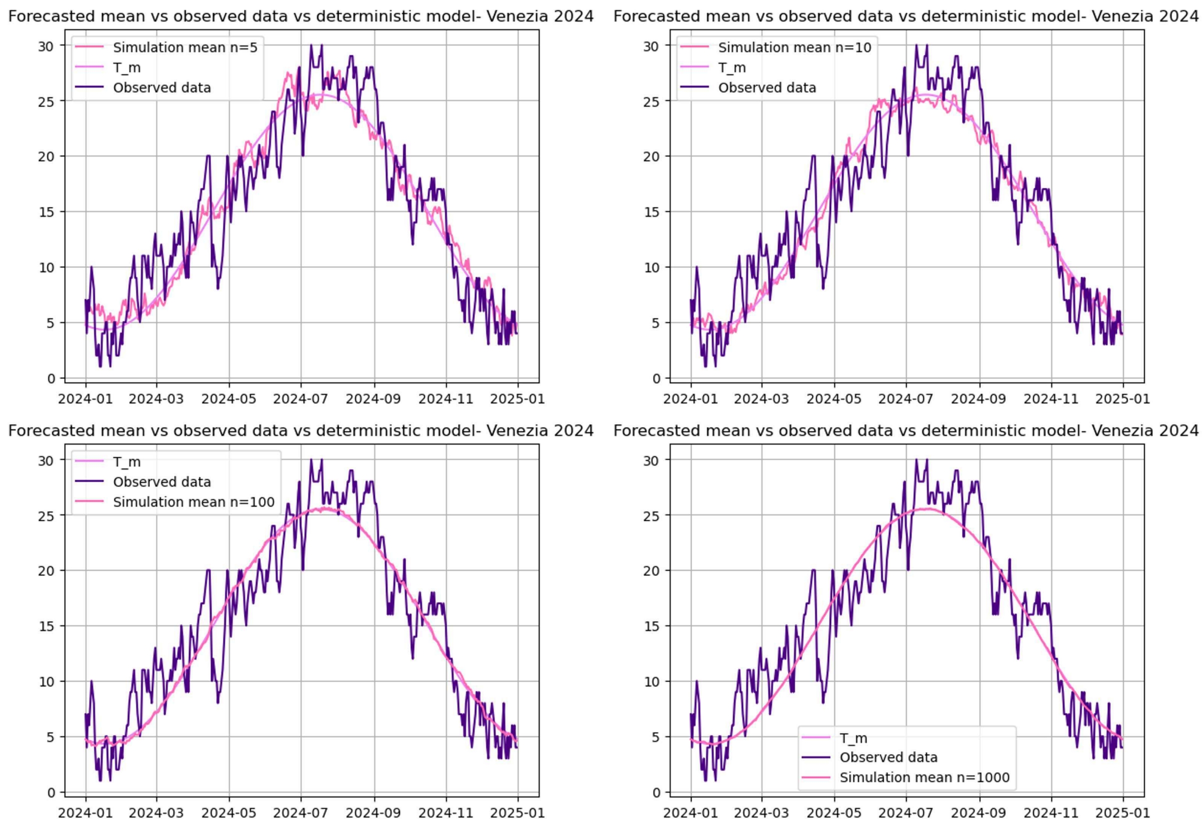


Figure 3.7: Illustration of the mean-reverting property of the process.

3.2.3 Volatility

While volatility is generally time-dependent in real-world temperature dynamics, for computational reasons we assume it to be piecewise constant. As discussed in Chapter 2, the empirical analysis shows that volatility does not remain constant over the entire year but can be reasonably approximated as constant within each calendar month. This assumption allows us to apply the known analytical solution of the Ornstein–Uhlenbeck process with constant parameters over each monthly interval, significantly simplifying the simulation of temperature trajectories.

Starting from the solution of the residual temperature process:

$$T_t = T_t^m + e^{-a(t-s)}(x - T_s^m) + \int_s^t \sigma_u e^{-a(t-u)}\, dW_u,$$

we simulate the evolution of T_t across multiple months by splitting the time interval $[s,t]$ into monthly subintervals over which the volatility σ is held fixed. For instance, to simulate temperature from January through March, the stochastic integral can be decomposed as:

$$\begin{aligned}\sigma \int_s^t e^{-a(t-u)}\, dW_u = \sigma_{Jan} \int_s^{\tau_1} e^{-a(t-u)}\, dW_u \quad &+ \sigma_{Feb} \int_{\tau_1}^{\tau_2} e^{-a(t-u)}\, dW_u \\ + \sigma_{Mar} \int_{\tau_2}^{t} e^{-a(t-u)}\, dW_u,&\end{aligned}$$

where $s =$ January 1, $\tau_1 =$ January 31, and $\tau_2 =$ February 28/29. Each of the three integrals can be evaluated by treating σ as constant within the corresponding month, making use of the closed-form distribution of the OU process under constant coefficients. In Table 3.4 are reported the estimated values computed as described in the Subsection 2.6.3.

Month	Bari	Viterbo	Rome	Venice	Turin	Bologna	Bolzano	Palermo	Ancona
1	1.873123	2.129885	1.870545	1.566466	2.241632	2.059548	1.669293	2.238333	1.711875
2	1.945820	2.134168	1.814852	1.778609	2.086932	1.805094	1.502150	2.162581	1.748819
3	1.981763	2.026564	1.908279	1.890395	2.104488	1.636931	1.389091	2.079728	2.037592
4	1.786481	1.883440	1.874780	1.817720	1.980902	1.521345	1.405703	2.120804	1.898942
5	1.871974	1.768298	1.894329	1.745871	2.029456	1.529935	1.577376	2.550092	1.918207
6	1.621496	1.650936	1.618780	1.512625	1.784025	1.301308	1.415689	2.206704	1.739293
7	1.767429	1.769987	1.533764	1.202578	1.773094	1.236839	1.442218	2.376119	1.773336
8	1.611022	1.770826	1.603996	1.288823	1.919680	1.399575	1.550580	2.213228	1.856541
9	1.651737	1.701681	1.489039	1.359275	1.847729	1.406351	1.503733	1.962374	1.803189
10	1.808765	1.673726	1.366977	1.573059	1.980523	1.633068	1.487942	1.848178	1.739787
11	1.946158	1.972553	1.489086	1.766945	2.069045	1.950502	1.585749	1.874743	1.569321
12	2.103192	2.394378	1.784495	1.750246	2.424940	2.354326	1.872480	2.252704	1.873869

Table 3.4: Monthly volatility estimates for each city.

3.2.4 Estimation of the mean-reversion parameter in the modified Ornstein-Uhlenbeck model

Due to the discrete nature of the available data, which consists of daily temperature observations, it is necessary to approximate the continuous-time model using a discrete-time formulation. For computational reasons, we have chosen a simple method based on the Euler-Maruyama discretization proposed by [12] of the stochastic differential equation, which allows for an efficient and straightforward estimation of the model parameters without requiring numerically intensive procedures.

Assuming a time step $\Delta t = 1$ (corresponding to one day), the discretized version of the process becomes:

$$T_{t+1} - T_t - T_{t+1}^m + T_t^m = a(T_m - T_t) + \sigma\varepsilon_t,$$

where the stochastic increments ε_t are modeled as independent and identically distributed standard normal random variables. Rearranging terms, we obtain a linear regression model of the form:

$$\Delta T_t = (a-1)(T_m - T_t) + \sigma\varepsilon_t,$$

where $\Delta T_t = T_{t+1} - T_t$ represents the daily variation in residual temperature, and $T_m - T_t$ is the deviation from the monthly mean.

In this formulation, the parameter a can be estimated via ordinary least squares (OLS) regression, treating ΔT_t as the dependent variable and $T_m - T_t$ as the independent variable. Since the parameter a may vary throughout the year due to seasonal effects, we perform the regression separately for each calendar month. That is, for each month $m = 1, \ldots, 12$, the model is fit using only the observations corresponding to that month, yielding a set of monthly estimates $\hat{a}_1, \ldots, \hat{a}_{12}$. A global estimate of the mean-reversion coefficient can then be computed as the average of the monthly estimates:

$$\bar{a} = \frac{1}{12}\sum_{m=1}^{12} \hat{a}_m.$$

This method offers several advantages. The estimation procedure is computationally efficient and easy to implement, making it well-suited for large datasets or iterative simulation procedures. The estimated parameter has a clear physical interpretation, representing the speed at which the system reverts toward the monthly mean. Moreover, estimating the coefficient separately for each month allows for the detection of potential seasonal patterns in the reversion behavior.

On the other hand, the method also presents some limitations. The Euler discretization introduces an approximation error with respect to the true transition density of the continuous-time process, and the estimation assumes that residuals are uncorrelated and homoscedastic, which may not hold in real temperature data. The monthly estimates of the mean-reversion parameter a were computed for each city in the dataset using the regression approach described above. Table 3.5 reports the estimated values of a for each month and city. These values reflect how quickly the residual temperature tends to revert to the monthly mean, with higher values corresponding to stronger reversion dynamics. A graphical representation of these estimates are shown in Figure 3.8.

Month	Bari	Viterbo	Rome	Ancona	Palermo	Bolzano	Turin	Bologna	Venice
1	0.254	0.313	0.262	0.294	0.263	0.309	0.255	0.207	0.191
2	0.232	0.228	0.188	0.230	0.262	0.339	0.196	0.164	0.168
3	0.255	0.257	0.229	0.210	0.342	0.217	0.190	0.182	0.131
4	0.264	0.220	0.216	0.257	0.370	0.205	0.188	0.207	0.150
5	0.269	0.189	0.171	0.226	0.308	0.332	0.211	0.211	0.171
6	0.162	0.149	0.124	0.196	0.261	0.254	0.169	0.177	0.148
7	0.244	0.207	0.177	0.272	0.291	0.322	0.206	0.241	0.196
8	0.286	0.234	0.211	0.284	0.334	0.287	0.207	0.227	0.197
9	0.239	0.275	0.192	0.241	0.284	0.243	0.203	0.233	0.208
10	0.281	0.265	0.221	0.246	0.320	0.225	0.145	0.180	0.170
11	0.314	0.339	0.308	0.324	0.421	0.223	0.214	0.194	0.205
12	0.314	0.336	0.306	0.312	0.312	0.317	0.213	0.198	0.220

Table 3.5: Monthly estimates of the mean-reversion parameter a for each city.

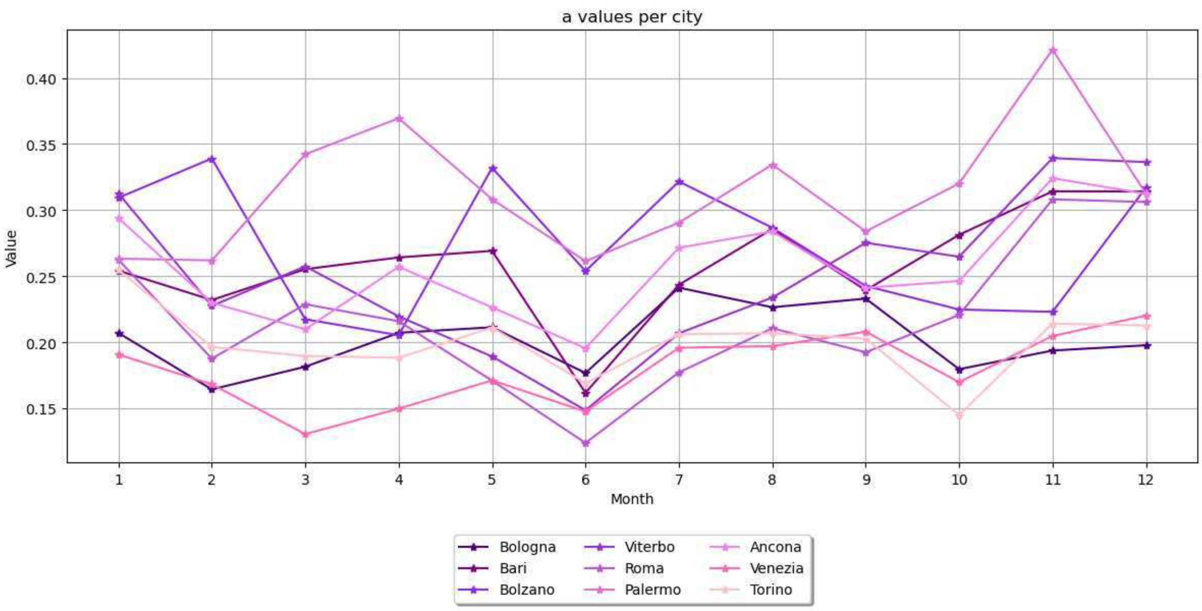


Figure 3.8: Monthly estimates of the mean-reversion parameter.

The estimated values show notable seasonal and geographical variation. Cities such as Palermo and Bolzano tend to exhibit relatively higher values of the mean-reversion parameter across multiple months, suggesting that temperature anomalies in these locations tend to dissipate more rapidly. In contrast, cities like Venice and Bologna show generally lower values, indicating a slower return to the climatological mean. These differences may reflect local climatic factors such as altitude, proximity to the sea, and continentality. Among the cities analyzed, Bolzano and Palermo show consistently higher values of the temperature inversion parameter a for several months. This behavior can be attributed to distinct climate mechanisms that favor a rapid decay of thermal anomalies. Bolzano is characterized by large diurnal temperature ranges and relatively low atmospheric humidity. These factors contribute to a rapid dissipation of thermal deviations, especially in clear-sky conditions and during the transition seasons. The limited thermal inertia of the local environment facilitates a stronger thermal inversion strength towards the mean. On the other hand, Palermo benefits from the stabilizing influence of the Mediterranean Sea. The maritime climate determines a lower intramonthly temperature variability and a tendency for anomalies to be short-lived. Furthermore, extreme thermal events, such as heat waves caused by Saharan winds, are typically transitory, with conditions quickly returning to normal. Despite their geographic and climatic differences, both cities have environmental characteristics that enhance the stability of the residual temperature process, resulting in higher estimates of the inversion parameter.

Figure 3.9 shows the behavior of the modified Ornstein-Uhlenbeck model applied to temperature data from 2005 to 2023 for the city of Venice, along with a zoomed-in view of the year 2015 to highlight the model's ability to capture short-term dynamics. In the plot on the right side of Figure 3.9, 20 simulated time series for the year 2015 generated by the model are shown, while the real observed data is highlighted in indigo.

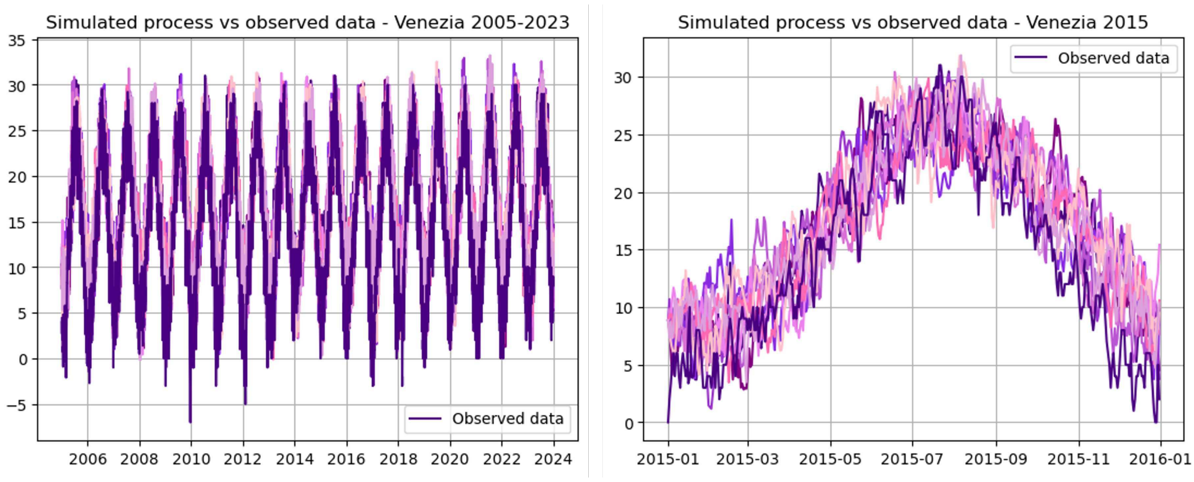


Figure 3.9: Simulated process for Venice and observed data.

The Figure 3.10 shows three visualizations of the model residuals: a time series plot, a histogram compared with the normal density and a QQ-plot. These plots provide a visual assessment of the distribution and characteristics of the residuals produced by the model.

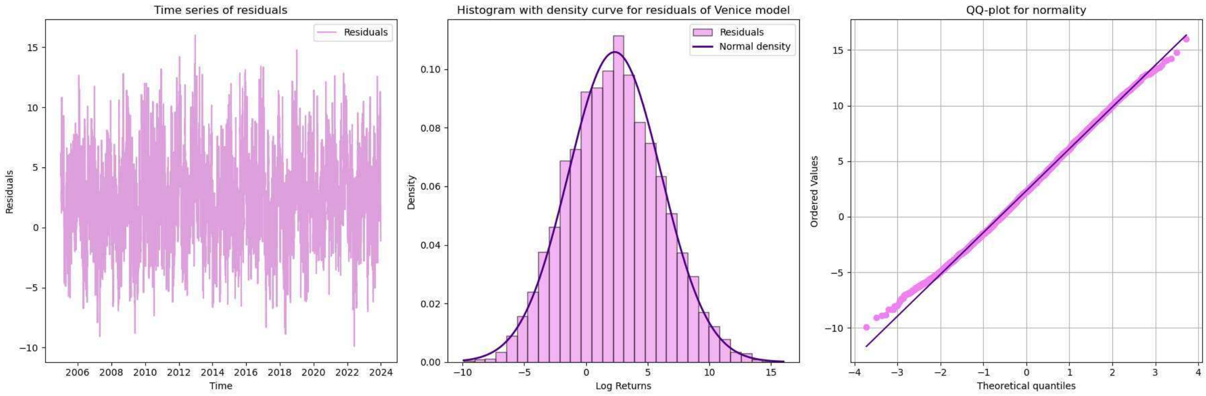


Figure 3.10: Residual of the Venice process.

3.2.5 Forecast for the 2024 year

In order to evaluate the predictive capacity of the estimated model, we continued by making a forecast for the year 2024, which was not included in the parameter calibration phase. This procedure allows us to test the behavior of the model on "out-of-sample" data, providing an accurate measure of the simulated dynamics. The choice of the year 2024 for the test is due to the presence of real data for the entire year, which would not have been possible if the choice had fallen on the year 2025. The simulated trend for 2024 is compared with the observed values, in order to highlight any structural discrepancies or significant deviations. This comparison also allows us to verify whether the model is able to correctly capture the average behavior and variability of the series in the forecast period. In order to rigorously evaluate the accuracy and reliability of the predicted temperature series, we decided to consider multiple evaluation metrics, each capable of capturing different aspects of the model quality in order to have a complete overview. This approach is particularly useful in the context of highly variable time series, such as daily temperatures, where traditional measures such as the Mean Absolute Error can be misleading, since they are strongly influenced by peaks or outliers. A model could reproduce the overall trend of the series well but still obtain a high MAE value due to a few extreme deviations. For this reason, it is useful to combine the MAE with other metrics including one that is more robust to the amplitude of local deviations, such as the Pearson correlation coefficient defined in [15].

Definition 3.2.2 (Mean Absolute Error)**.** *The Mean Absolute Error is defined as the average of the absolute differences between the observed values y_i and the predicted values $\hat{y}_i$, i.e.:*

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n is the number of observations.

The Pearson correlation coefficient measures the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 no linear correlation. In this context, a high positive Pearson correlation indicates that the simulated and real series tend to move together, reproducing similar dynamics, even if the exact daily values may differ. This metric is particularly appropriate in contexts where the goal is to capture the temporal behavior and structure of the series rather than to provide precise point-wise prediction. In fact, error-based metrics such as RMSE or MAE can be misleading in highly volatile settings, as small temporal misalignments or local fluctuations may result in disproportionately high penalties. Pearson's correlation, on the contrary, provides a robust summary of the model's ability to reproduce the dynamics of the underlying process.

Definition 3.2.3 (Pearson correlation)**.** *We can define the Pearson correlation as follows*

$$r = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}} \tag{3.9}$$

where x_t and y_t represent the values of the real and simulated series at time t, and $\bar{x}_t$ and $\bar{y}_t$ are their respective sample means. This formula computes the normalized covariance between the two series, making it invariant to scale and location shifts.

The last metrics we used is the Mean Bias Error.

Definition 3.2.4 (Mean Bias Error)**.** *The Mean Bias Error is the average of the differences between the observed values y_i and the predicted values $\hat{y}_i$. It is defined as:*

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

where n is the number of observations.

A positive MBE indicates that the model is underestimating the values on average, meaning predictions tend to be lower than actual values. A negative MBE suggests that the model is overestimating the values, with predictions being higher than actual ones. An MBE close to zero implies that the model has little or no bias, meaning the predictions are, on average, close to the observed values. The closer the MBE is to zero, the better the model's calibration. The combined use of these three metrics therefore allows a more robust and balanced evaluation of the model:

- the MAE measures the average absolute accuracy,
- the Pearson correlation evaluates the structural similarity between real and simulated series,
- the MBE quantifies the presence of directional biases.

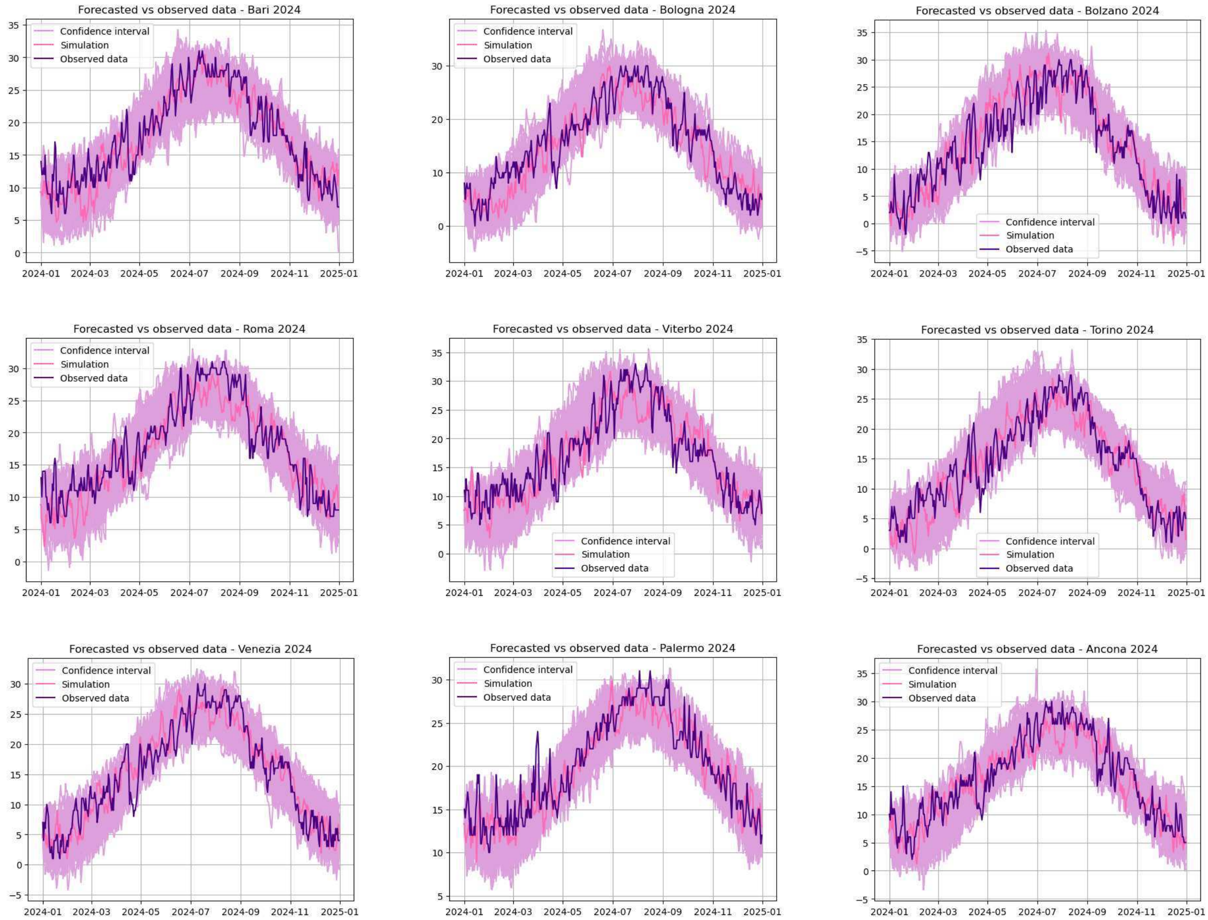


Figure 3.11: Forecast with Model 1 for 2024.

	Palermo	Bari	Venezia	Bolzano	Torino	Roma	Ancona	Bologna	Viterbo
MAE	1.829	2.324	2.330	3.073	2.675	2.731	2.710	2.945	2.892
MBE	-0.368	-0.417	0.112	0.481	-0.281	-0.752	-0.583	-0.619	-0.140
Pearson	0.911	0.909	0.932	0.906	0.904	0.879	0.887	0.897	0.879

Table 3.6: Values of MAE, MBE, and Pearson correlation coefficient for each city based on the comparison between the simulated and actual 2024 temperature series.

The Table 3.6 summarizes the metrics with which we want to compare the simulated and real temperature series for the year 2024. The MBE values are generally low and close to zero, indicating that the models do not have an evident systematic bias, i.e. they do not tend to constantly overestimate or underestimate temperatures. The Pearson correlation, high in all cities (always above 0.87), shows that the general dynamics of the simulated data faithfully follow the trend of the observed temperatures. Although the MAE values are on average between 2.5 and 3 degrees, it is important to consider that these are forecasts on an annual time horizon of a notoriously volatile variable such as temperature. Furthermore, a tendency is observed according to which coastal cities, such as Palermo and Bari, obtain better results, with lower absolute errors and a slightly higher correlation. On the contrary, for the inland cities, such as Bologna, Bolzano and

Turin, the performances tend to worsen, probably due to the greater complexity of the local thermal dynamics, more sensitive to orographic and continental factors. These observations support the idea that the model is able to adequately capture the general trends, although showing some difficulties in the most internal and variable areas.

3.2.6 Impact of historical window length on model parameters and performance

In this section we analyze how the choice of the width of the time window used for calibration affects the forecasting capabilities of the modified Ornstein–Uhlenbeck model. As for the deterministic model, three different historical horizons have been considered: 5, 10 and 19 years, with the 19-year model used as a reference for comparison since it is the one used to calibrate and forecast throughout the chapter. Particular attention is paid to the mean reversion parameter (a) whose annual mean values are reported in Table 3.7. The estimated values of the parameter show good stability across the different time windows in many cities, where variations are generally less than 2%. However, more marked deviations are observed in cities such as Palermo and Bolzano, in particular in the case of the 5-year calibration. This suggests that these locations have undergone recent changes in the mean level of the process, which are captured more sensitively by models based on shorter time series. It is also noteworthy that this change occurred in the two cities that had a higher reference a parameter than the others. From a forecasting perspective, the

City	**Bari**	**Viterbo**	**Roma**	**Ancona**	**Palermo**	**Bolzano**	**Torino**	**Bologna**	**Venezia**
5 years	0.2599	0.2514	0.2053	0.2560	0.2733	0.2737	0.1981	0.1996	0.1801
10 years	0.2539	0.2514	0.2171	0.2558	0.3279	0.2645	0.2012	0.2025	0.1758
19 years	0.2596	0.2510	0.2171	0.2577	0.3142	0.2728	0.1998	0.2017	0.1795

Table 3.7: Comparison of parameter a values across time horizons for each city.

5-year calibrated models, whose metrics results are reported in Table 3.9, tend to generate lower MAE values in some cases like Palermo. These results highlight how the use of recent data favors short-term forecasting, as it allows to better capture the current dynamics and recent trends. In particular, it was seen that for Palermo there was a significant change in the mean reversion parameters that could lead us to suspect recent climate changes in the area that would explain the better performance of the model trained only with more recent data. Despite the gain in some cases, this can be accompanied by bias as only very recent data has been used which is not capable of capturing long trend behaviour. This reflects the risk of overfitting to transitory fluctuations or local anomalies in the most recent data. The 10-year models show a more balanced profile. In terms of MAE, they often equal or slightly exceed the 19-year model, while maintaining low bias and high correlation coefficients. Pearson values, always above 0.86 in all configurations, indicate

a strong linear coherence between forecasts and observed data, regardless of the length of the calibration window. Putting it all together, while 5-year models are very reactive to the changes of recent years in the trends, they tend to be too sensitive to the same recent fluctuations. The 10-year model, on the other hand, could be a fair compromise as it ensures solid performance and great parameter stability. 19-year models, although less reactive, are preferable for long-term analyses, thanks to their capability to capture longer trends. At the computational level, the use of models calibrated on 5- or 10-year windows is certainly more efficient, thanks to the significant reduction in data required.
Remark. However, considering that the differences between the 10-year model and the full 19-year model are limited, one could reasonably opt for the former. Nevertheless, in order to ensure greater robustness and reliability of the estimates, it was decided to keep, in the subsequent analyses, the entire historical series available as a training window for the model under examination.

	Bologna	**Bari**	**Bolzano**	**Viterbo**	**Roma**	**Palermo**	**Ancona**	**Venezia**	**Torino**
MAE	3.0047	2.8421	3.2635	2.8679	2.8004	1.8738	2.8660	3.0083	2.7283
MBE	-0.1588	-1.4150	-0.5135	-0.3692	-0.2968	-0.3672	-0.7365	-0.2865	0.3862
Pearson	0.8962	0.8823	0.8881	0.8833	0.8683	0.9022	0.8819	0.8825	0.9014

Table 3.8: Performance metrics across cities - 10 years model.

	Bologna	**Bari**	**Bolzano**	**Viterbo**	**Roma**	**Palermo**	**Ancona**	**Venezia**	**Torino**
MAE	2.9771	2.5983	3.3383	3.0883	2.7671	1.7710	2.8811	2.9645	2.7133
MBE	0.7484	0.1177	0.3441	0.2043	-0.7266	-0.1314	-0.4866	-0.0401	1.0940
Pearson	0.8879	0.8885	0.8950	0.8712	0.8875	0.9116	0.8742	0.8942	0.9004

Table 3.9: Performance metrics across cities - 5 years model.

3.3 Model 2: ARMA and GARCH

Once the deterministic component of the process is isolated, expressed in the same way as the previous model with a linear trend and a seasonality marked by a sine function, the residuals obtained represent the stochastic part of the process. These residuals do not appear purely random in fact they present autocorrelation and a variance that seems to vary over time. For this reason, it was decided to model the residual dynamics using a combination of ARMA and GARCH models in order to capture the linear dependencies in the residuals and flexibly describe the variability of the conditioned variance. In the literature on modeling temperature time series, a wide range of autoregressive models have been proposed to capture both the deterministic and stochastic components of climatic data. Traditional ARMA and ARIMA models have served as foundational tools to describe short-term temporal dependencies and are frequently applied after the removal of trends and seasonal components. However, standard ARMA models alone often fail to capture the persistent long-memory behavior observed in climatic variables so ARCH and GARCH models have been adapted to climate time series to model time-varying conditional variance. Another extension of the ARMA is found in the ARFIMA models that allows to effectively capturing the long-memory behavior often observed in temperature time series by [4]. This makes ARFIMA particularly suitable for modeling persistent climatic phenomena where autocorrelations decay slowly over time. Continuous-time models such as CARMA have also been applied to capture complex temporal dependencies; for example, F. E. Benth in [2] employs a CAR(3) process to model temperature dynamics. However, given that the available data are recorded at daily intervals, a discrete-time modeling approach was preferred for this study.

3.3.1 Preliminary exploration of residues

Temperature modeling using an ARMA+GARCH approach requires preliminary verification of some fundamental assumptions about the process residuals: in particular, it is necessary to ensure that these residuals are stationary on average and that they exhibit conditional heteroskedasticity. Stationarity is a fundamental concept in time series analysis and is of great importance when applying models such as ARMA (Autoregressive Moving Average). Stationarity implies that the statistical properties of a process, such as its mean, variance and autocovariance, are independent of time. In other words, to correctly apply models like the one we are going to define, it is necessary that the process is stationary, otherwise the forecast results may be biased or inconsistent. In particular, a stationary process does not exhibit long-term trends and its fluctuations are regular over time, which makes it suitable for models that capture serial dependence and conditional variability.

Definition 3.3.1. *A weakly stationary process X_t has the following main characteristics:*

- *For all t, $\mathbb{E}[X_t] = \mu$ with μ a constant;*
- *For all t, $Var(X_t) = \sigma^2$ with σ a constant;*
- *Constant auto-covariance between the values of the process at a fixed distance (lag) i.e. $Cov(X_t, X_{t+k}) = \gamma(k)$.*

It is therefore essential to remove non-stationary components, such as long-term trends or cycles, before applying an ARMA model because this class of models are based on the assumption that the future value of a time series depends on its past values and a random error. However, this relationship is only valid if the process is stationary. If a series has a deterministic trend, such as linear growth, the ARMA model may mistakenly try to model this trend as well, distorting the forecasts. Therefore, before applying an ARMA model, it is necessary to differencing the time series to remove non-stationary components and ensure that the series is stationary. In case the series is not stationary, appropriate transformations, such as differencing, should be applied to obtain a stationary process before applying the ARMA model. In the our case the trasformation is alredy done differienting the observed value with the deterministic model that removes the trends and the seasonal cycles. To be completly sure, it was decided to implement a test to verify whether or not the stochastic part of the series is stationary. The most widely used tool in this context for testing stationarity is the Augmented Dickey-Fuller (ADF) test. The ADF test is designed to test for the presence of a unit root in a time series, which would be indicative of a non-stationary component in the process. Before introducing the meaning of a unit root as defined in [19], it is essential to begin with the definition of the lag operator and the representation of autoregressive processes. The lag operator (L) is a mathematical convention used to express relationships between current and past values of a variable. Applying L to a variable Y_t shifts it back by one period, such that $LY_t = Y_{t-1}$, and generally $L^k Y_t = Y_{t-k}$. For example the first difference $\Delta Y_t = Y_t - Y_{t-1}$ can be compactly written as $(1 - L)Y_t$. Consider an autoregressive process of order p where the current value Y_t linearly depends on its p past values and a white noise error term ϵ_t: $Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$. By employing the lag operator, this equation can be rewritten as:

$$Y_t - \phi_1 L Y_t - \phi_2 L^2 Y_t - \ldots - \phi_p L^p Y_t = c + \epsilon_t.$$

Factoring out Y_t, we obtain:

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) Y_t = c + \epsilon_t.$$

The term in parentheses is a polynomial in the lag operator, commonly denoted as

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p,$$

allowing the AR(p) process to be expressed as $\Phi(L)Y_t = c + \epsilon_t$. To ascertain the stationarity properties of an AR(p) process, one must analyze the roots of the characteristic equation associated with the polynomial $\Phi(L)$. This characteristic equation is formed by substituting the lag operator L with a complex variable (typically z or λ) and setting the resulting polynomial equal to zero: $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$. The roots of this equation are the values of z that satisfy it. A fundamental condition for an AR(p) process to be stationary is that all roots of its characteristic equation must have a modulus strictly greater than one. A unit root occurs when at least one of the roots of the characteristic equation is precisely equal to one in modulus. The simplest illustration is the random walk ($Y_t = Y_{t-1} + \epsilon_t$), where the characteristic equation $1 - z = 0$ yields a root

$z = 1$. To statistically assess the presence of such a root, the Augmented Dickey-Fuller (ADF) test is employed. In this context, the null hypothesis of the ADF test is that the time series possesses a unit root, i.e., it is non-stationary. Rejection of the null hypothesis thus provides evidence in favor of stationarity. In this study, the test is conducted at the standard 5% significance level.

Definition 3.3.2. *The ADF test considers the following linear regression:*

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \varepsilon_t$$

where:

- Y_t *is the value of the time series at time* t*,*
- $\Delta Y_t = Y_t - Y_{t-1}$ *is the first difference of the series,*
- α *is a constant,*
- βt *is a trend term,*
- γ *is the autoregressive coefficient,*
- ε_t *is the random error.*

The test checks whether the coefficient γ *is equal to zero. If* γ *is significantly different from zero, the series is stationary.*

To assess the stationarity of the residual time series, we applied the ADF test to each city's series. The test evaluates the null hypothesis that a unit root is present, implying non-stationarity. As shown in Table 3.10, all p-values are well below the conventional threshold of 0.05, allowing us to reject the null hypothesis in every case. These results confirm that the residual series are stationary. This finding is consistent with what can be observed in the Figure 3.12.

It has already been seen that these series are stationary and it is of particular interest to delve deeper by analyzing whether or not any autocorrelation structures are present that should be modeled with an ARMA process. For this purpose, it is possible to use the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) that allow identifying the presence of linear dependence between the values of the series.

Definition 3.3.3 (ACF)**.** *The autocorrelation function of order* k *is defined as:*

$$\rho(k) = \frac{Cov(X_t, X_{t-k})}{Var(X_t)} = \frac{\mathbb{E}[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2}$$

City	ADF Statistic	p-value	Conclusion
Bologna	-21.9172	0.0000	Stationary
Bari	-18.9807	0.0000	Stationary
Bolzano	-19.0062	0.0000	Stationary
Viterbo	-18.4154	0.0000	Stationary
Roma	-20.7663	0.0000	Stationary
Palermo	-10.7033	0.0000	Stationary
Ancona	-20.5577	0.0000	Stationary
Venezia	-22.1484	0.0000	Stationary
Torino	-19.2972	0.0000	Stationary

Table 3.10: ADF test results for each city's temperature residuals.

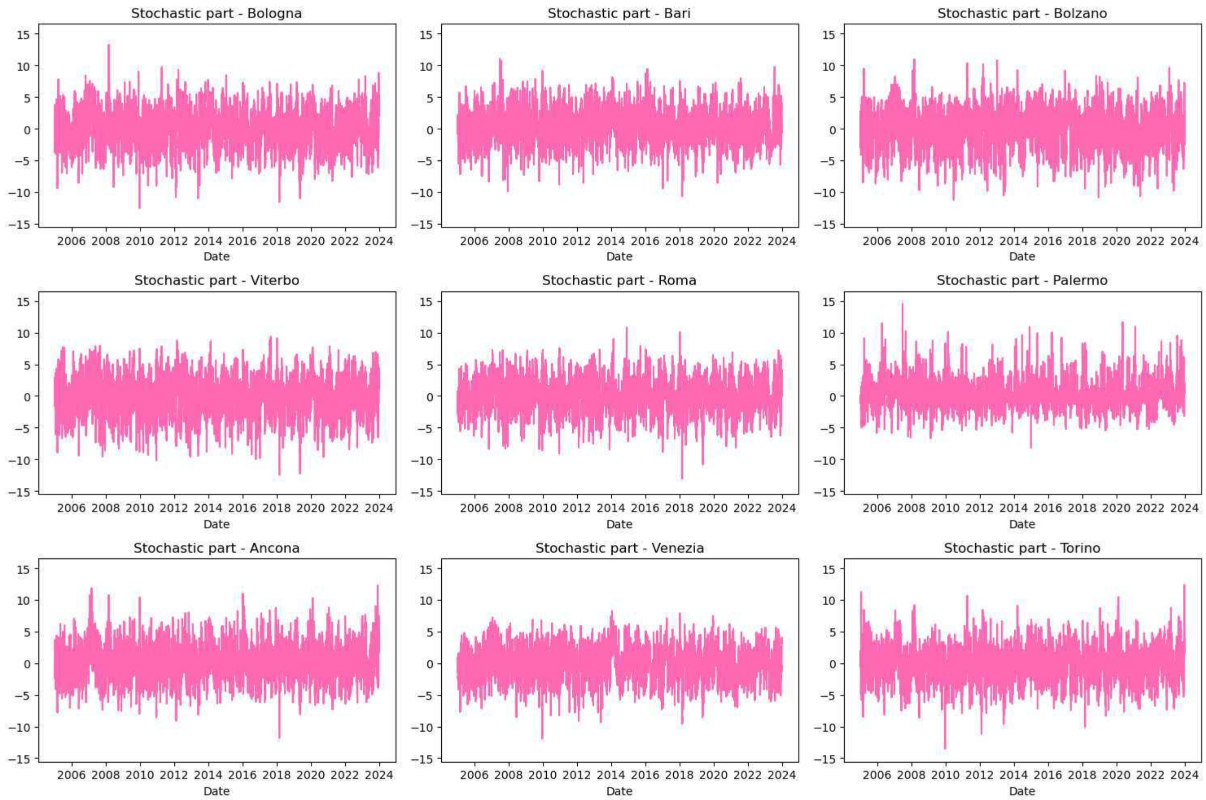


Figure 3.12: Temperature residuals after removing the deterministic component.

where μ is the mean of the series (assumed constant for stationarity) and σ^2 its variance. The ACF therefore measures the linear correlation between X_t and X_{t-k} at various lags k, and is useful for identifying the MA(q) *component of the model.*

Definition 3.3.4 (PACF)**.** *The partial autocorrelation function (PACF) of order k, instead, measures the correlation between X_t and X_{t-k}* removing the effect of all intermediate lags $1, 2, \ldots, k-1$. *It is obtained through successive regressions or by solving the Yule-Walker equations. The PACF is useful for identifying the* AR(p) *component of the model.*

Looking at the ACF and PACF graphs in figure 3.13, we notice clear peaks at certain lags suggesting the need to define a model that takes these dependencies into account. This type of behavior justifies the adoption of an $ARMA(p, q)$ model, which allows to effectively capture the autoregressive and moving average dynamics of the residuals, thus improving the predictive capacity of the overall model.

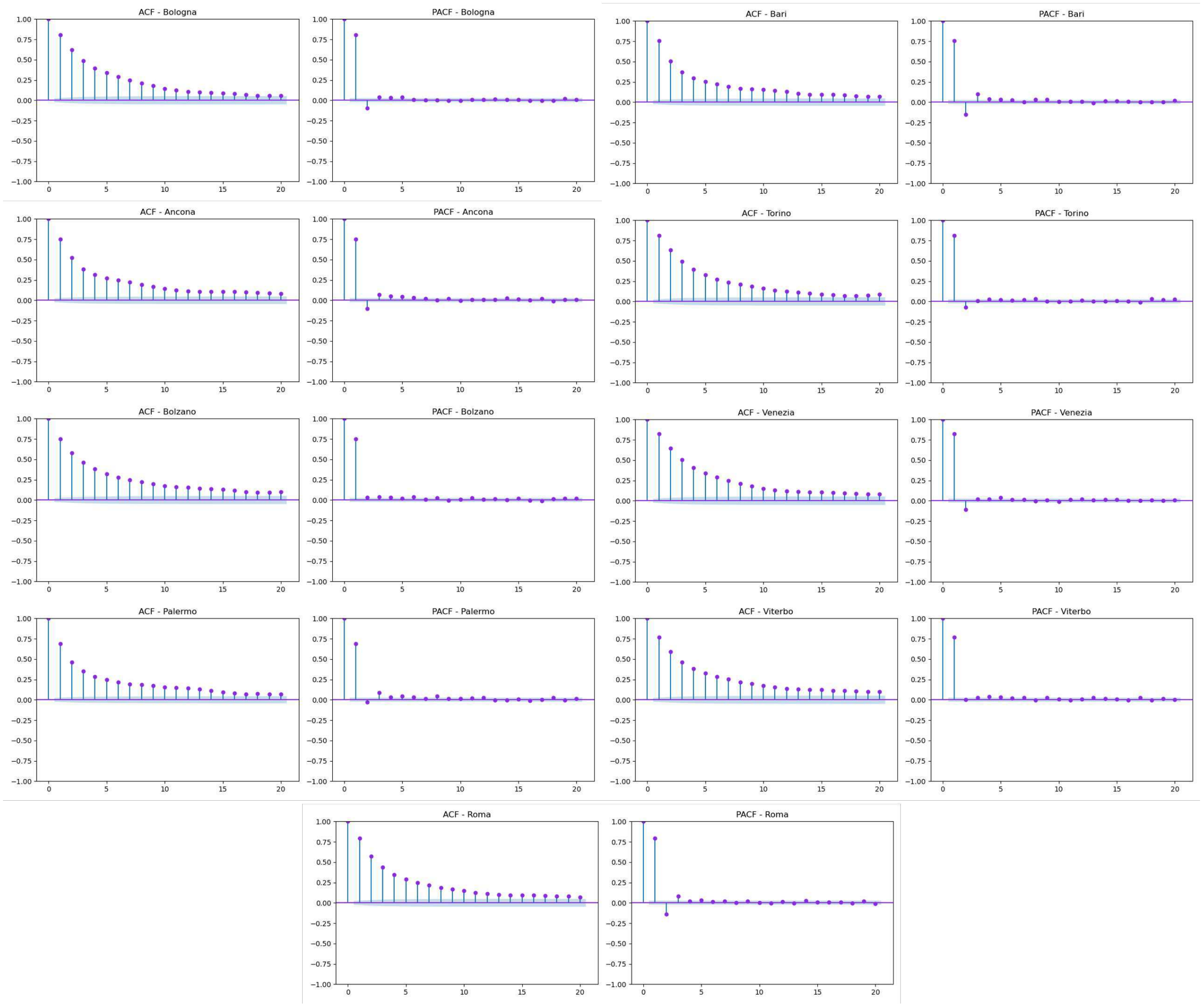


Figure 3.13: ACF and PACF of the temperature residuals.

3.3.2 ARMA

From the previous analysis, it was observed that the residual temperature series exhibit both stationarity and significant autocorrelation. These characteristics motivate the adoption of an Autoregressive Moving Average model to adequately capture the linear temporal dependencies present in the data. Before continuing, some definitions taken from [19].

Definition 3.3.5 (White Noise). *W is called a white noise series if it satisfies the following conditions:*

- *For all t, $\mathbb{E}[W_t] = \mu$ with μ a costant;*

- *For all t,* $Var(W_t) = \sigma^2$ *with* σ *a positive costant;*
- *It is uncorrelated at different time points: when* $t \neq s$, $Cov(W_t, W_s) = 0$.

Definition 3.3.6 (ARMA model). *The following equation is called the autoregressive moving average model of order* (p, q) *and denoted bu* $ARMA(p, q)$*:*

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{3.10}$$

where

- X_t *it is the time series value at time* t*;*
- ϕ_i *autoregressive coefficients (AR);*
- θ_j *moving average coefficients (MA);*
- ε_t *it is a white noise process with 0 mean, costant variance* σ^2 *and zero serial correlation:* $E[\varepsilon_t] = 0$, $Var(\varepsilon_t) = \sigma^2$ *and* $Cov(\varepsilon_t, \varepsilon_{t-k}) = 0 \;\; \forall k \neq 0$.

If X *is weakly stationary and satisfies the equation (3.10), then we call it an* $ARMA(p, q)$ *process.*

In the process, the AR component captures the linear dependence between the current observation and the past values, while the MA component captures the dependence of the current value on the past innovations.

Model identification

The first crucial step is to identify the correct structure of the process. This requires determining the orders p (autoregressive) and q (moving average). This process can be conducted through a joint approach that combines graphical analysis and informative criteria.

- The Autocorrelation Function (ACF) provides information on the linear dependencies between an observation and its lags. For a pure MA(q) process, the ACF shows significantly different values from zero up to lag q, and then it goes to zero.
- The Partial Autocorrelation Function measures the correlation between X_t and X_{t-k} by removing the effect of intermediate lags. For a pure AR(p) process, the PACF stops after the delay p.

Joint observation of ACF and PACF allows to formulate preliminary hypotheses on the possible (p, q) combinations. However, visual identification can be supported by quantitative criteria, in order to automate and optimize the model selection.

To this end, we use the Akaike Information Criterion (AIC), defined in [19] as:

$$\text{AIC} = -2\log(L) + 2k,$$

where L is the value of the likelihood function of the estimated model and k represents the total number of free parameters. The model with the lowest AIC is preferred, as it provides a good compromise between the quality of the fit and the parametric complexity.

In practice, automatic model selection functions are often used, such as `auto_arima()` available in the Python `pmdarima` library. This algorithm systematically explores combinations of orders (p, q) within a specified range and selects the optimal model by minimizing the AIC (or BIC if preferred). The `auto_arima()` procedure was used to identify the optimal ARMA(p, q) structure for each series. This method was preferred due to its automated nature, as it systematically explores different combinations of models and selects the one that minimizes the AIC. This ensures objective and efficient model selection, especially when dealing with multiple time series.

In Table 3.11 are reported the values of p and q for each city, estimated using the `auto_arima()` function. The estimation was carried out on the daily temperature time series from 2005 to 2023. The resulting orders are generally low, which is a desirable property: parsimonious models are typically more robust, easier to interpret, and less prone to overfitting, especially when forecasting. This observation aligns with findings by [2], who showed that for Swedish temperature data, even an ARMA model with parameters $p = 3$ and $q = 0$ was sufficient to capture the main temporal dynamics.

	Bologna	**Bari**	**Bolzano**	**Viterbo**	**Roma**	**Palermo**	**Ancona**	**Venezia**	**Torino**
p	3	2	2	1	2	2	3	1	2
q	1	2	1	0	2	3	1	1	3

Table 3.11: Parameters p and q.

Estimation of the parameters The next step is the estimation of the ARMA parameters which was carried out using the classically used method in these cases of maximum likelihood. The statistical procedure allows the identification of estimates for the consistent coefficients. The Table 3.12 reports the estimated values for the models of the individual cities using the historical series from 2005 to 2023. The values with an asterisk next to them represent those statistically insignificant with a p value less than 0.05. The values, instead, with two asterisks are statistically insignificant if the chosen confidence interval is 99%. The choice made is to maintain a confidence interval of 95%. In this way the only non-significant parameters are the constant terms for each city.

	const	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3	σ^2
Bari	0.000306*	1.499935	-0.520648	–	-0.625592	-0.256410	–	2.837969
Palermo	0.000632*	1.549185	-0.568330	–	-0.844102	-0.123334	0.063195	2.169139
Bologna	0.001802*	1.588426	-0.755195	0.112994	-0.700383	–	–	2.909079
Ancona	0.002514*	1.608938	-0.795785	0.143730	-0.784920	–	–	3.143860
Torino	0.001498*	1.641166	-0.657240	–	-0.772130	-0.090748	−0.035825**	2.572040
Roma	0.000282*	1.530853	-0.559949	–	-0.625090	-0.211089	–	2.457905
Viterbo	0.000070*	0.767733	–	–	–	–	–	3.624819
Venezia	0.000075*	0.781395	–	–	0.136048	–	–	2.120968
Bolzano	−0.002538*	1.621208	-0.638107	–	-0.902162	–	–	4.068713

Table 3.12: Estimated ARMA parameters: * p-value>0.05, ** p-value>0.01.

3.3.3 GARCH

In ARMA models one of the assumptions is that the variance of the residuals is constant over time. However, this assumption is often violated with time series. In fact, it has already been seen previously that volatility required at least a division into monthly values. Real data very often show periods of calm followed by others of greater turbulence. This phenomenon is known as conditional heteroscedasticity and implies that the variance of the errors depends on past values and therefore is not constant over time. To model this phenomenon, a first solution was provided in 1982 by Engle with the Autoregressive Conditional Heteroscedasticity (ARCH) model, which models the variance as a function of time. Before continuing, it is useful to provide some definitions taken from [19].

Definition 3.3.7 (ARCH). *An ARCH(m) model with order $m > 0$ is of the form*

$$\begin{cases} X_t = \sigma_t \varepsilon_t \\ \sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \alpha_2 X_{t-2}^2 + \cdots + \alpha_m X_{t-m}^2 \end{cases}$$

where $\omega \geq 0$, $\alpha_i \geq 0$ and $\alpha_m > 0$ are costants, $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and ε_t is independent of $\{X_k : k \leq t-1\}$.

The model for a variance as a function of time was further improved by adding a dependence on past observations of the variance itself in addition to the dependence on the past values of the shocks. This new model called Generalized Autoregressive Conditional Heteroscedasticity (GARCH) was introduced by Tim Bollerslev in 1986.

Definition 3.3.8 (GARCH). *A GARCH(m,s) model with the order $m \geq 1$ and $s \geq 0$ is of the form*

$$\begin{cases} X_t = \sigma_t \varepsilon_t \\ \sigma_t^2 = \omega + \sum_{k=1}^m \alpha_k \varepsilon_{t-k}^2 + \sum_{\ell=1}^s \beta_\ell \sigma_{t-\ell}^2 \end{cases}$$

where $\omega \geq 0$, $\alpha_i \geq 0$,$\beta_i \geq 0$, $\alpha_m > 0$ and $\beta_s > 0$ are costants, $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and ε_t is independent of $\{X_k : k \leq t-1\}$.

To understand whether or not it is necessary to use in our time series in addition to an ARMA model also the GARCH for volatility, two approaches were followed. The first is graphical with the display of the residuals of the ARMA model squared. Reported in the Figure 3.14, it can be seen how ACF and PACF show in all cases the presence of significant lags indicating the possible presence of heteroscedasticity. To confirm the hypothesis made by observing the plots, the ARCH-LM statistical test was performed. The test involves regressing the squared residuals $\hat{\varepsilon}_t^2$ from the initial model on q of their own lagged values. Specifically, the auxiliary regression is given by:

$$\hat{\varepsilon}_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \hat{\varepsilon}_{t-i}^2 + u_t$$

From this regression, the coefficient of determination R^2 is obtained, and the test statistic is computed as $\mathrm{LM} = TR^2$, where T denotes the number of observations. Under the null hypothesis of homoskedasticity (i.e., no ARCH effects), the test statistic asymptotically follows a chi-squared distribution with q degrees of freedom:

$$\mathrm{LM} \sim \chi_q^2$$

If the null hypothesis is rejected, this indicates the presence of ARCH effects in the residuals, suggesting that a GARCH-type model is suitable to model the conditional variance. The results for the ARCH-LM test are reported in Table 3.13 while the ACF and PACF of the squared residuals are in Figure 3.14

City	LM Stat	LM p-value	F Stat	Interpretation
Bari	184.23	3.09e-34	18.90	ARCH effects detected
Palermo	368.74	4.18e-73	38.88	ARCH effects detected
Bologna	57.39	1.13e-08	5.78	ARCH effects detected
Ancona	101.16	3.19e-17	10.25	ARCH effects detected
Torino	65.68	3.01e-10	6.62	ARCH effects detected
Roma	201.25	8.84e-38	20.69	ARCH effects detected
Viterbo	89.71	6.12e-15	9.07	ARCH effects detected
Venezia	41.20	1.04e-05	4.14	ARCH effects detected
Bolzano	40.36	1.47e-05	4.05	ARCH effects detected

Table 3.13: ARCH LM test results for each city. Interpretation is based on LM p-values (< 0.05).

Both the plots of the squared residuals and the results of the statistical tests show the presence of heteroskedasticity in line with what was observed in the previous chapter where the analysis of the volatility of temperatures showed a non-constant temporal structure that could be best approximated on a monthly basis. From these observations it seems appropriate to continue modeling the residuals through a GARCH model capable of

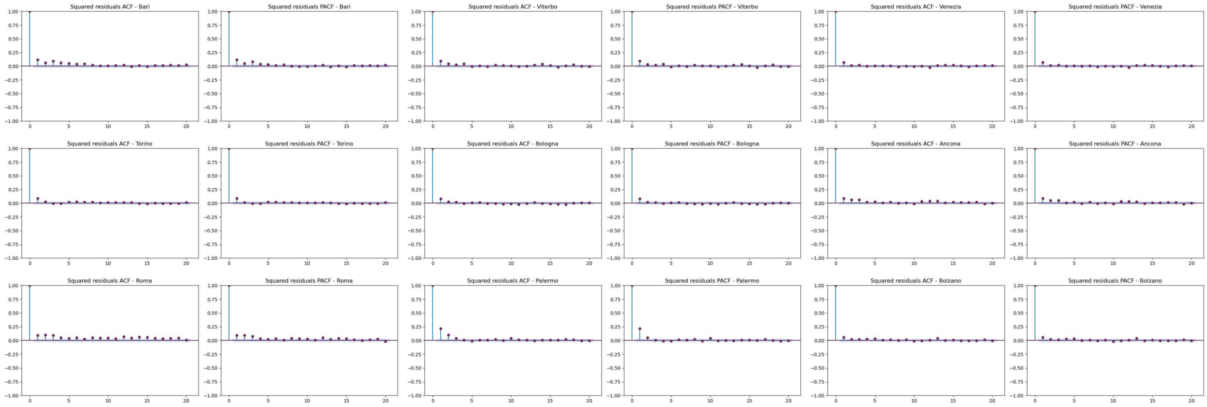


Figure 3.14: ACF and PACF of squared residuals.

capturing the variations in variance over time. For the distribution of innovations ε_t it was decided to use a Student-t instead of the normal one as it is more suitable for managing heavy tails in the data. This choice is supported by the fact that in the historical series there may be extreme events such as cold or heat waves that could generate significant deviations. A GARCH(1,1) model was applied to all series and was found to be suitable in all situations to capture the dynamics of the conditional variance. The parameters, estimated with maximum likelihood assuming a Student-t distribution for innovations, are reported in the Table 3.14. In Figure 3.15, the trend of the conditional volatility, the ACF of the standardized residuals, the histogram and the QQ-plot are shown for each series, useful for evaluating the adequacy of the modeling and the distribution of the post-GARCH errors.

Parameter	Bari	Palermo	Bologna	Ancona	Torino	Roma	Viterbo	Venezia	Bolzano
Omega	0.6731	0.9303	1.7029	0.8225	0.0615*	0.0604	1.3833	1.3998	1.4998*
Alpha1	0.1495	0.3291	0.0764	0.1000	0.0175	0.0547	0.1033	0.0850	0.0644
Beta1	0.6201	0.3106	0.3385	0.6407	0.9585	0.9220	0.5193	0.2599	0.5697
Nu	9.6425	4.0673	8.6027	8.6430	8.7900	7.4702	7.8832	9.0713	7.1819

Table 3.14: Estimated parameters for GARCH(1,1)-t: * p-value $>$0.05.

At the level of results we can say that we are satisfied. The residuals do not show further signs of correlation indicating that all the information present in the series has been correctly modeled. As for the distribution of the residuals we can note how most of the cities follow a Student-t quite well. Turin and Bologna present a slight negative asymmetry while Palermo has a longer right tail than expected.

3.3.4 Forecast for the 2024 year

As already done for the first model, to evaluate the predictive capabilities we continued by making the forecast for the entire year 2024. Therefore, for each series, after having estimated an ARMA model for the mean and GARCH for the conditional variance, we

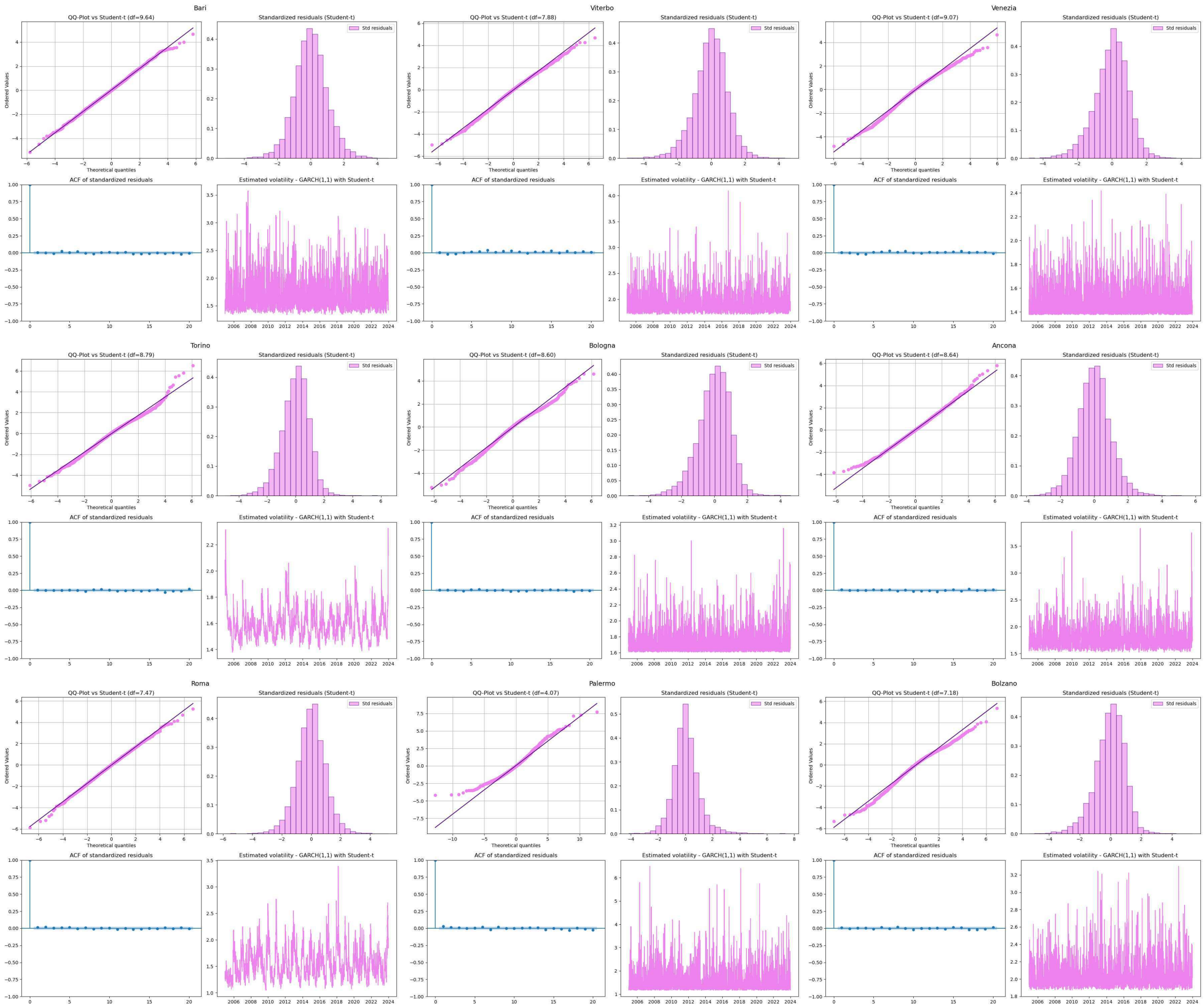


Figure 3.15: Graphical results of the model for each city: conditional volatility, ACF of residuals, histogram and QQ-plot.

Metric	Bari	Palermo	Bologna	Ancona	Torino	Roma	Viterbo	Venezia	Bolzano
MAE	2.321229	2.004335	2.737429	2.724086	3.247285	2.703204	3.064374	2.491425	3.408989
MBE	0.057468	0.572611	0.392865	0.635731	-0.426446	-0.138111	0.140389	0.388603	0.305517
Pearson	0.900010	0.889164	0.907238	0.885080	0.870479	0.875559	0.869746	0.920061	0.873514

Table 3.15: Values of MAE, MBE and Pearson correlation coefficient for each city.

moved on to implement a forecasting process. The objective was to obtain an estimate of future temperatures. These forecasts were carried out with a recursive procedure, present in the Appendix, which involved the use of the previously predicted values for the subsequent steps. This choice allowed us to avoid that the variance and the mean converged towards the constant parameters that are zero in all cases, in order to maintain a realistic representation of the temperatures.

To evaluate the predictive capacity of the model, the simulated temperatures for the year 2024 were compared with the real ones we had available. In order to have a comparison with the first proposed model, the metrics chosen for the evaluation are the same as those used previously. Table 3.15 reports the values for the Mean Absolute Error, Mean Bias Error and for the Pearson correlation coefficient. It can be observed that once again the coastal cities are the best represented by the Model 2 as was observed in Model 1, with the values for Bari, Palermo and Venice decidedly lower than those of inland cities such as Bolzano and Viterbo. The correlation remains however remarkably high and once again we do not find extremely overestimated or underestimated temperatures. It should be noted that the first proposed model performs slightly better than this one to which is added the computational complexity clearly lower than the first.

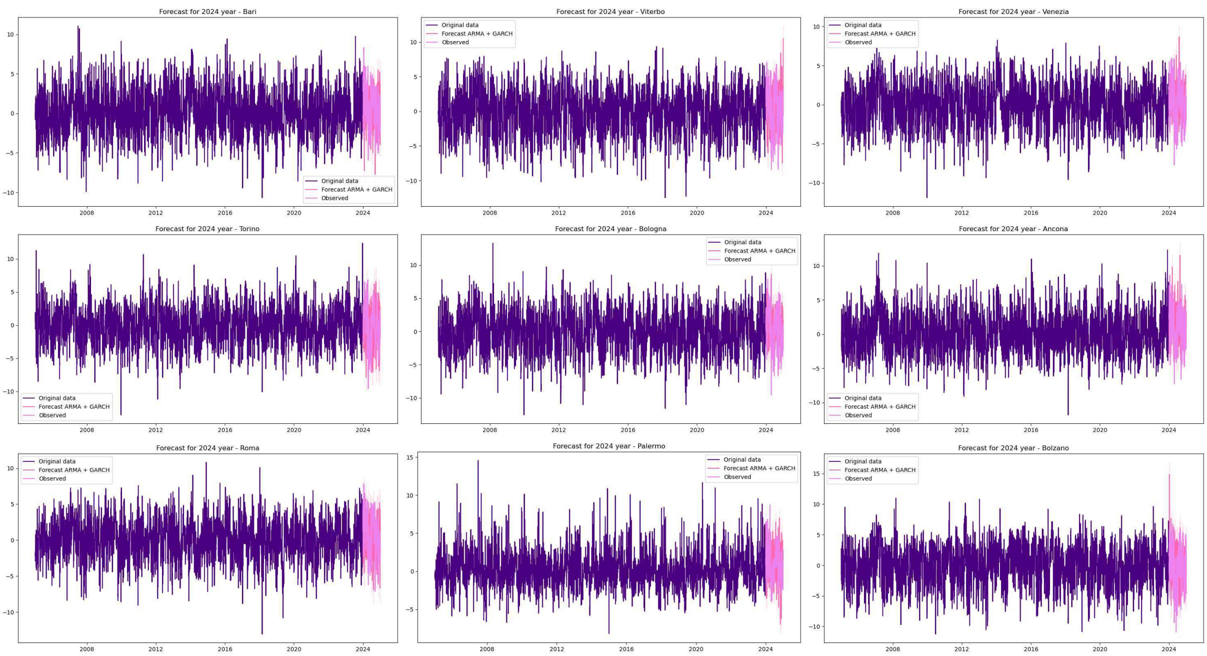


Figure 3.16: Forecast results for the stochastic part for the 2024 year.

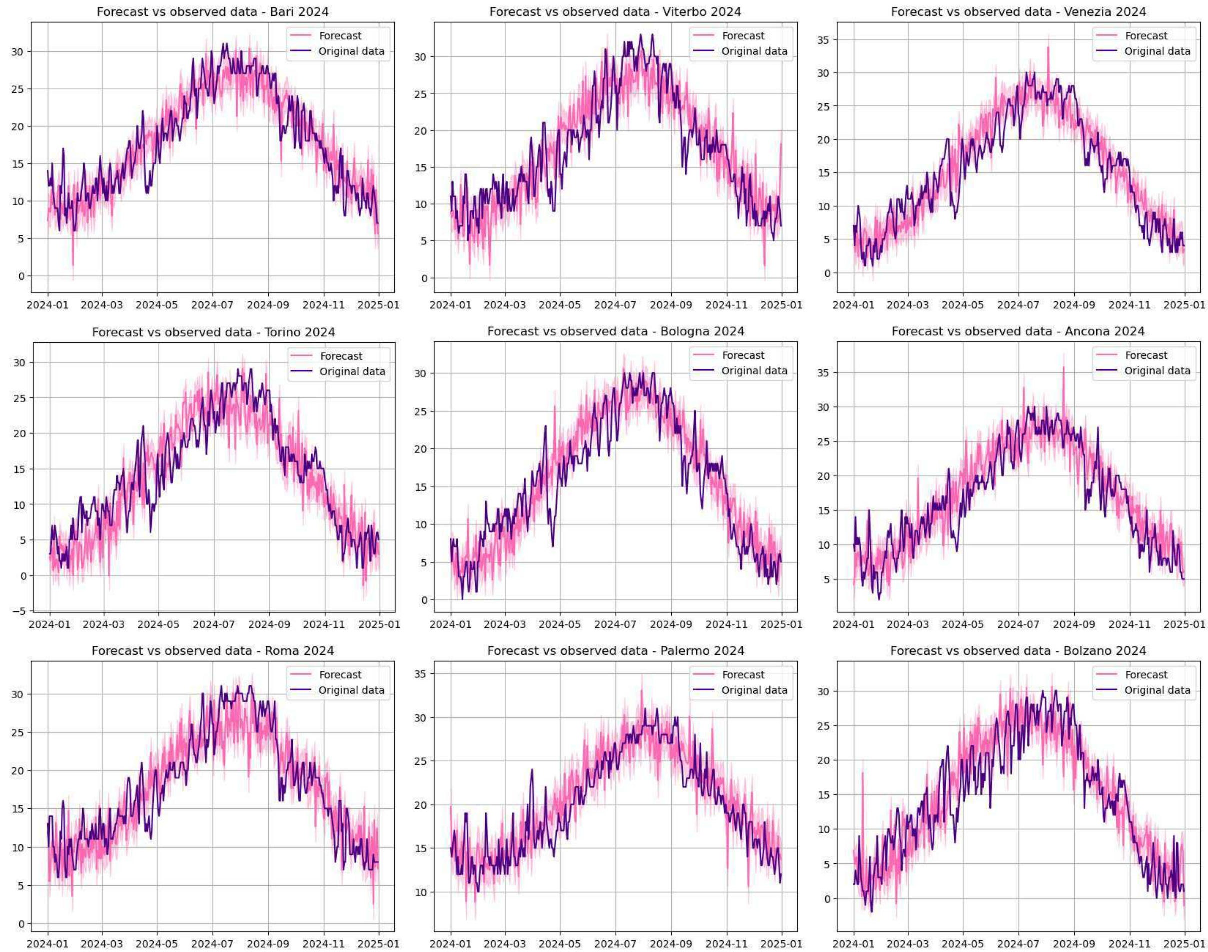


Figure 3.17: Forecast results for the 2024 year.

The Model 2 was applied to the residual component of the temperature series, obtained after removing a deterministic function accounting for trend and seasonality. The forecasts were produced on this stochastic component, and subsequently, the deterministic part was reintroduced to reconstruct the full temperature forecast. Figure 3.16 shows the forecasted trajectories of the stochastic component for the year 2024, while Figure 3.17 displays the corresponding full forecasts after adding back the deterministic component, allowing for a direct comparison with the observed data. The Model 2 shows good potential in capturing the seasonality and general dynamics of temporal data for 2024 however, some discrepancies in peaks and variability suggest room for improvement, especially in the ability to follow rapid and local variations. Overall, the model provides a reliable forecast but with a level of uncertainty that varies depending on the city and the period considered.

3.3.5 Impact of historical window length on model parameters and performance

As was done for the previous model, we will analyze how the length of the time series used to calibrate the model affects the predictive capabilities of the model. The approach followed is the same with the choice of the same time windows of 5, 10 and 19 years. The latter is the reference one since everything presented up to this point used all the available data that are equivalent to 19 years. Since the deterministic component has already been discussed previously, in this section we will focus only on the ARMA and GARCH parameters. The first thing to analyze is if and how the parameters p and q change with the different series. A comparison is reported in Table 3.16. It can be observed that in some cities such as Bari, Rome and Bolzano and Venice the values for the autoregressive and moving average components remain identical all three times indicating good stability. The cities of Turin and Palermo report slight changes probably due to events present in the time series. The particular case is Viterbo which reports a high instability in the parameters probably due to climate changes in the area.

Parameter	Model	Bologna	Bari	Bolzano	Viterbo	Roma	Palermo	Ancona	Venezia	Torino
	19 years	3	2	2	1	2	2	3	1	2
p	10 years	3	2	2	3	2	2	2	1	3
	5 years	2	2	2	2	2	2	3	1	2
	19 years	1	2	1	0	2	3	1	1	3
q	10 years	1	2	1	1	2	2	2	1	2
	5 years	2	2	1	1	2	2	1	1	2

Table 3.16: p and q parameters for the different length of the time-series used for training.

Looking at the table 3.17 in which the values of the estimated parameters for the individual cities in all three models are reported with the different lengths of the series used for training, it is possible to see how for most of the cities the parameters remain almost similar. The substantial differences are visible in the cities of Viterbo where the greatest difference compared to the other cities is due to the changes in the number of parameters in the ARMA.

Parameter	Model	Bari	Palermo	Bologna	Ancona	Torino	Roma	Viterbo	Venezia	Bolzano
const	5 years	-0.006485	-0.004918	-0.002459	0.011128	0.003984	0.003601	0.006996	0.004181	0.004357
	10 years	0.013761	0.004402	0.008691	0.010574	0.002955	0.007372	0.010642	0.003345	-0.004715
	19 years	0.000306	0.000632	0.001802	0.002514	0.001498	0.000282	0.000070	0.000075	-0.002538
ϕ_1	5 years	1.382176	1.300197	1.457918	1.609833	1.667853	1.405764	-0.201317	0.747808	1.529845
	10 years	1.476887	1.378365	1.743775	1.471314	2.041087	1.520835	1.618550	0.779005	1.566373
	19 years	1.499935	1.549185	1.588426	1.608938	1.641166	1.530853	0.767733	0.781395	1.621208
ϕ_2	5 years	-0.422845	-0.347782	-0.499621	-0.872141	-0.681631	-0.449619	0.714706	–	-0.558266
	10 years	-0.495920	-0.409057	-0.879391	-0.494687	-1.340095	-0.543976	-0.661455	–	-0.586490
	19 years	-0.520648	-0.568330	-0.755195	-0.795785	-0.657240	-0.559949	–	–	-0.638107
ϕ_3	5 years	–	–	–	0.213192	–	–	–	–	–
	10 years	–	–	0.116999	–	0.287593	–	0.025393	–	–
	19 years	–	–	0.112994	0.143730	–	–	–	–	–
θ_1	5 years	-0.577269	-0.659015	-0.628039	-0.766276	-0.847219	-0.543472	0.961999	0.111911	-0.856310
	10 years	-0.636683	-0.707966	-0.885892	-0.660538	-1.190768	-0.630592	-0.893846	0.109234	-0.885469
	19 years	-0.625592	-0.844102	-0.700383	-0.784920	-0.772130	-0.625090	–	0.136048	-0.902162
θ_2	5 years	-0.214346	-0.109619	-0.153632	–	-0.069808	-0.197610	–	–	–
	10 years	-0.241178	-0.134162	–	-0.212994	0.261225	-0.226674	–	–	–
	19 years	-0.256410	-0.123334	–	–	-0.090748	-0.211089	–	–	–
θ_3	5 years	–	–	–	–	–	–	–	–	–
	10 years	–	–	–	–	–	–	–	–	–
	19 years	–	0.063195	–	–	-0.035825	–	–	–	–
σ^2	5 years	2.665316	2.228804	3.132474	3.132205	2.712352	2.372716	3.663918	2.113398	4.233813
	10 years	2.757369	2.184257	2.931147	3.141095	2.576981	2.495502	3.566452	2.106840	4.254735
	19 years	2.837969	2.169139	2.909079	3.143860	2.572040	2.457905	3.624819	2.120968	4.068713
ω	5 years	0.5771	0.8782	1.9033	0.2383	0.0412	0.1424	3.3614	1.8240	0.5864
	10 years	1.7002	0.6482	0.9016	1.8471	0.0930	0.9363	0.2092	1.3961	0.0643
	19 years	1.7029	0.6731	1.4998*	1.3833	0.0604	0.9303	0.8225	1.3998	0.0615*
α_1	5 years	0.0877	0.1718	0.0478	0.0923	0.0670	0.3111	0.0341	0.0600	0.0155
	10 years	0.0669	0.1446	0.0703	0.1071	0.0584	0.3102	0.0353	0.0949	0.0168
	19 years	0.0764	0.1495	0.0644	0.1033	0.0547	0.3291	0.1000	0.0850	0.0175
β_1	5 years	0.3031	0.6202	0.8137	0.0000	0.8752	0.3667	0.8907	0.0776	0.9690
	10 years	0.3522	0.6259	0.7201	0.3813	0.9057	0.3225	0.8982	0.2445	0.9584
	19 years	0.3385	0.6201	0.5697	0.5193	0.9220	0.3106	0.6407	0.2599	0.9585
ν	5 years	8.2783	14.2546	8.9636	7.5959	6.5794	3.9729	6.5458	12.3703	8.5609
	10 years	8.9778	12.8092	7.4383	8.1977	6.7553	4.0286	7.1753	10.7037	9.4334
	19 years	8.6027	9.6425	7.1819	7.8832	7.4702	4.0673	8.6430	9.0713	8.7900

Table 3.17: ARMA and GARCH parameters for the different time windows (5, 10, 19 years).

A comparison of the performances of the three models with different training time horizons, 5, 10 and 19 years, was carried out by observing the three metrics used so far: the mean absolute error (MAE), the mean bias error (MBE) and the Pearson correlation coefficient. These durations represent the length of the temperature time series used to train the models. As shown in Table 3.18, it can be observed how, in general, the use of a shorter time window leads to a worsening of the MAE values. This highlights how a time series of only 5 years may be insufficient to capture the real trend of temperatures. With 10 years of data the situation improves significantly, but the performances remain lower than those of the model based on 19 years of data. As for the MBE, the values are more variable between the different models and cities, without indicating a systematic tendency

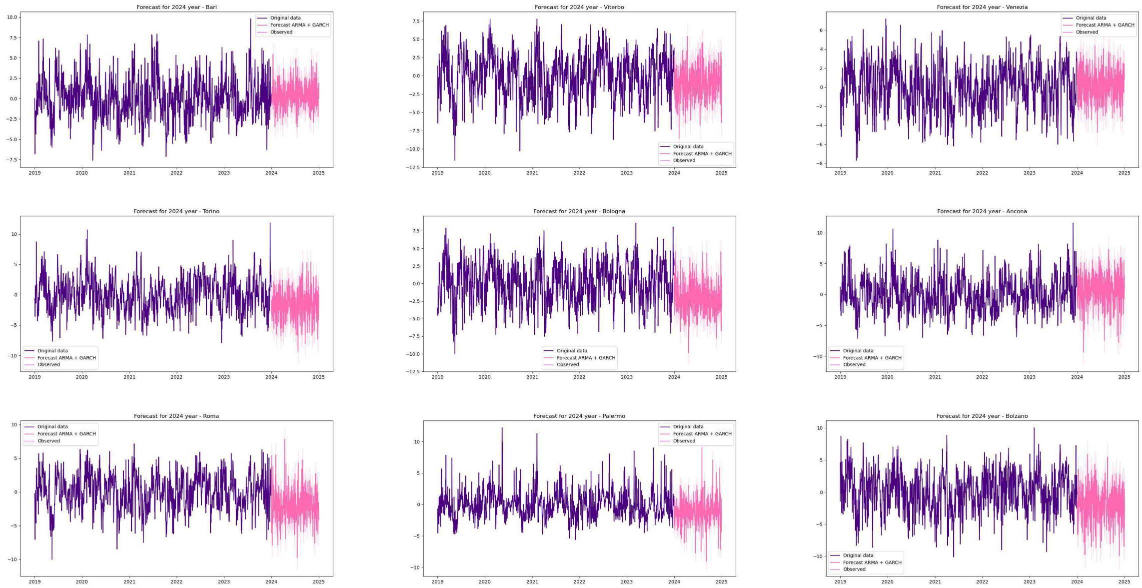


Figure 3.18: Forecast results for the stochastic part with 5 years time-series lenght.

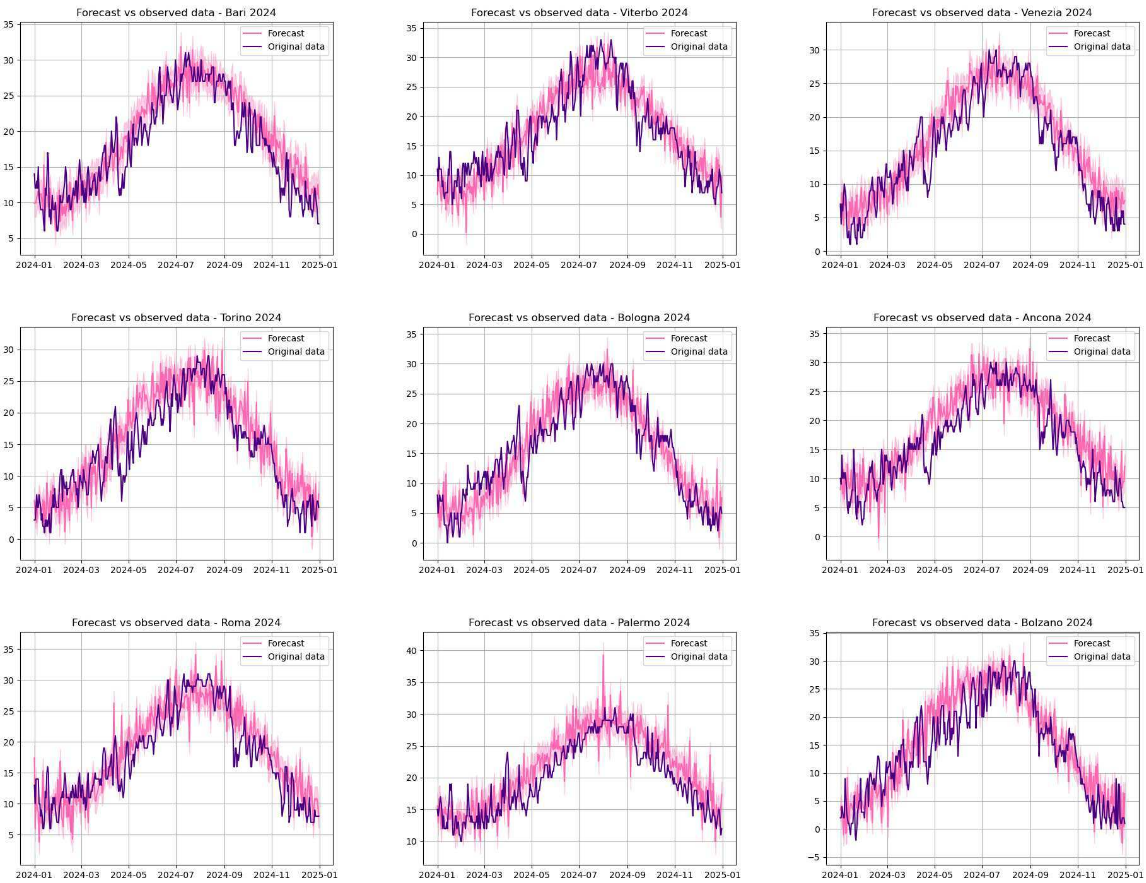


Figure 3.19: Forecast results with 5 years time-series lenght.

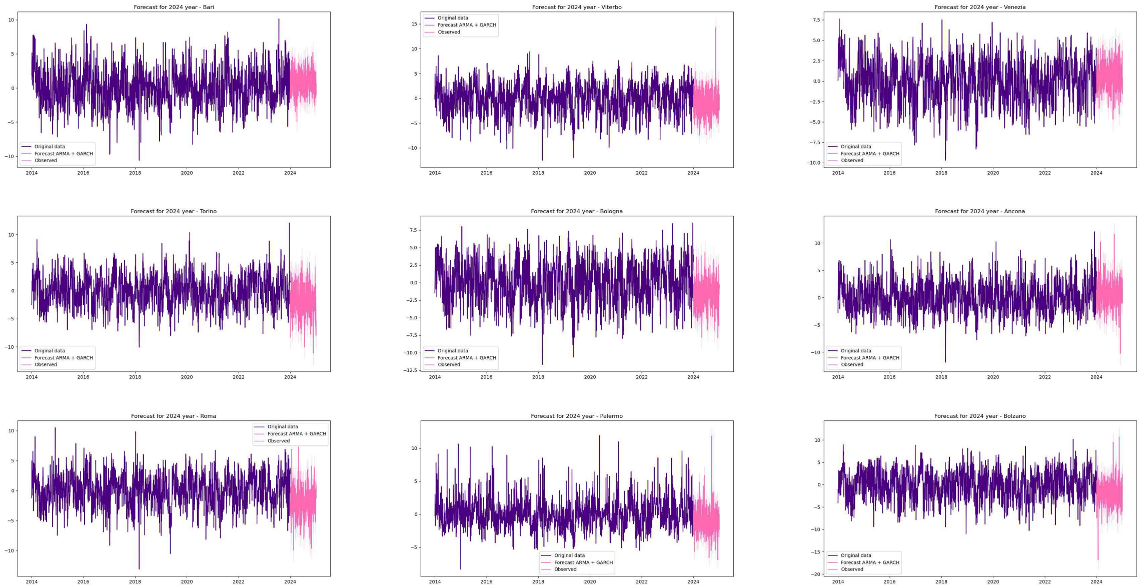


Figure 3.20: Forecast results for the stochastic part with 10 years time-series lenght.

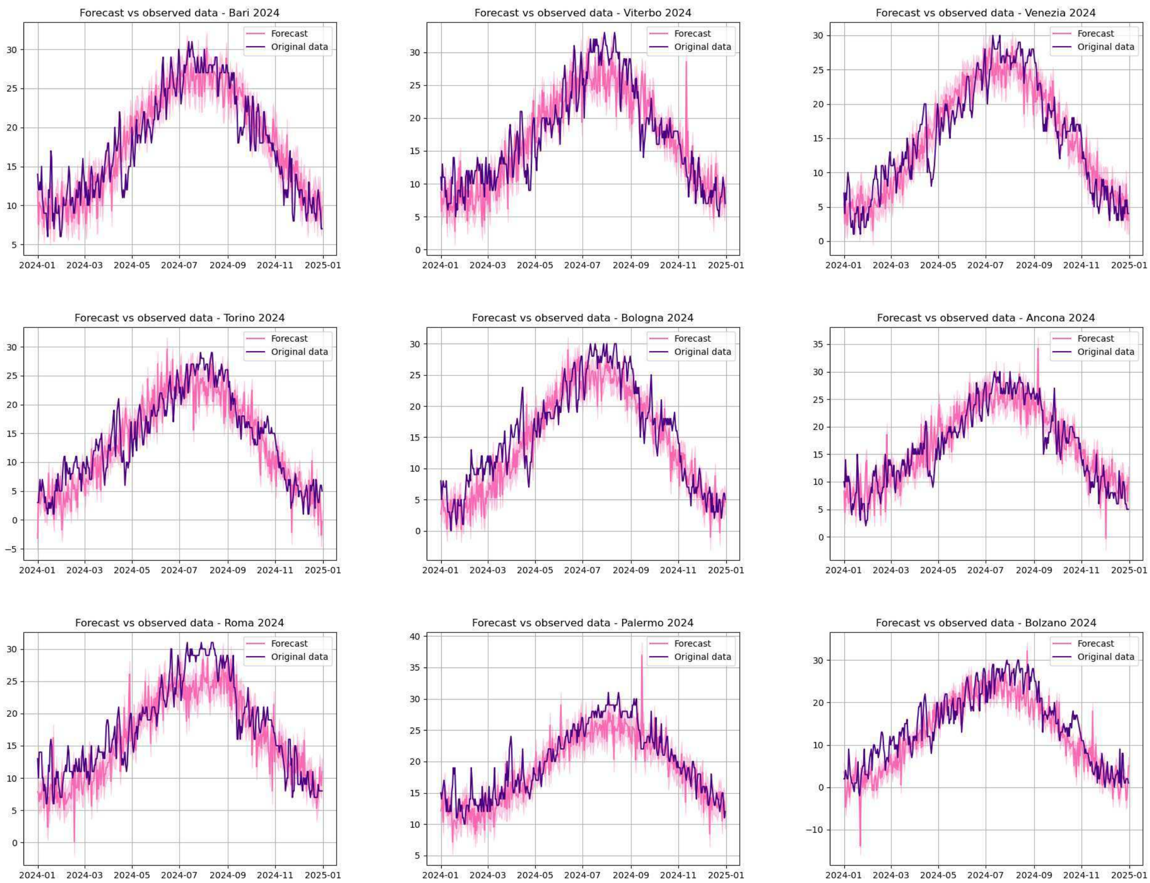


Figure 3.21: Forecast results with 10 years time-series lenght.

to overestimate or underestimate the observed values. The Pearson correlation coefficient, however, remains quite stable in all cases, showing a constant ability of the models to follow the temporal trend of the series. Figures 3.19, 3.18, 3.20 and 3.21 show in detail the comparisons between the deterministic plus stochastic component and the stochastic component alone for the 5- and 10-year models. In conclusion, the results confirm once again that the use of the complete historical series, equal to 19 years, represents the best choice in terms of forecasting model performance.

Metric	**Horizon**	**Bari**	**Palermo**	**Bologna**	**Ancona**	**Torino**	**Roma**	**Viterbo**	**Venezia**	**Bolzano**
MAE	5 years	2.4718	2.5513	2.9396	3.1846	3.1952	2.9007	3.0396	2.6217	3.3404
	10 years	2.3380	2.2521	3.1296	2.7258	3.0294	3.2388	3.0856	2.6372	3.8068
	19 years	2.3212	2.0043	2.7374	2.7241	3.2473	2.7032	3.0644	2.4914	3.4090
MBE	5 years	1.1392	1.6301	0.1617	1.7781	1.3331	0.8720	0.1385	1.1899	0.9543
	10 years	0.1103	-1.2086	-1.3263	0.2302	-0.7331	-1.4404	-0.4330	-0.0817	-2.0647
	19 years	0.0575	0.5726	0.3929	0.6357	-0.4264	-0.1381	0.1404	0.3886	0.3055
Pearson	5 years	0.9053	0.8829	0.9003	0.8692	0.8790	0.8763	0.8710	0.9227	0.8838
	10 years	0.9031	0.8743	0.9022	0.8774	0.8779	0.8533	0.8606	0.9164	0.8763
	19 years	0.9000	0.8892	0.9072	0.8851	0.8705	0.8756	0.8697	0.9201	0.8735

Table 3.18: MAE, MBE and Pearson metrics for each forecast horizon (5, 10, and 19 years) across cities.

3.4 Comparative analysis of predictive models

With the aim of comparing the performance of the two models proposed in this chapter, a comparative analysis is presented based on the three metrics considered in the evaluation of the individual models. Overall, Model 1 (modified Ornstein-Uhlenbeck) obtained better performances in terms of absolute value as it is lower in six cities out of the nine analyzed. Furthermore, the Pearson correlation values are generally higher for the OU model, suggesting greater coherence in following the temporal evolution of the target variable. Another point in favor of this first model can be found in the fact that it is less expensive at the computational level. Although the Model 2 (ARMA+GARCH) is theoretically more suitable for modeling dynamics with heteroskedasticity, in this specific context the modified OU model shows greater stability and generalization on cities with heterogeneous climate and behavioral profiles. These results suggest that, at least for the application domain considered, the OU model represents a more solid methodological choice. To further explore the differences between the two approaches, specific observations for each city are provided below, with reference to the three metrics considered.

- Palermo: The Model 1 is clearly superior in all metrics. It shows the best MAE of the group.
- Bari: Very similar performances for both models, with an advantage of the Model 2 in MBE.

- Venice: Model 1 offers higher performance in all metrics. Model 2 shows marked positive bias.
- Bolzano: Model 1 has a slightly lower MAE and a higher Pearson but the Model 2 shows a smaller MBE.
- Turin: Model 1 shows a better performance.
- Rome: Model 1 shows more negative MBE but maintains a comparable level of MAE. Similar Pearson between models.
- Ancona: Both models have similar MAE and Pearson correlation but they differ for the MBE with Model 1 that shows a marked negative bias and Model 2 a marked postive one.
- Bologna: Model 2 shows a lower MAE with a lower bias then Model 1.
- Viterbo: Model 1 performs better in all metrics, with lower MAE and MBE and a higher Pearson.

Figure 3.22: Comparison of Model 1 and Model 2 across cities based on MAE, MBE and Pearson correlation metrics.

Chapter 4

Pricing temperature derivatives

"It is better to be vaguely right than exactly wrong."
— Carveth Read

Over the past few decades, the growth of climate unpredictability has led to the emergence of financial instruments to hedge against weather risks. Among these financial instruments, temperature derivatives are among the most widespread. These contracts are typically based on cumulative temperature indices such as Heating Degree Days, Cooling Degree Days and Cumulative Average Temperatures, already introduced in Chapter 1. Unlike traditional derivatives, temperature derivatives and climate derivatives in general are not traded in highly liquid markets. The market for these indices is not only illiquid but also incomplete due to the impossibility of trading the underlying and storing it. Therefore, pricing weather derivatives requires an alternative approach based on probabilistic modeling of the underlying temperature process. Since there is generally no liquid market from which to extract implicit probabilities or risk-neutral dynamics, traders often resort to historical simulation methods with real-world probability measures. In this chapter, the aim is to present a theoretical framework for pricing temperature derivatives. Next, we examine the specific structures of the most common temperature indices (HDD and CDD) and their payoffs. Finally, we discuss the implementation of this pricing methodology using Monte Carlo simulations, leveraging the stochastic models for temperature dynamics developed in Chapter 3.

4.1 Equivalent measure and market price of risk

In classical financial theory, derivatives are priced based on the existence of an equivalent martingale measure where by the discounted price of the underlying asset becomes a martingale. However, this cannot be directly applied to weather derivatives for two main reasons:

- The underlying variable is not tradable. Temperature indices such as HDD or CDD are not financial assets and cannot be bought or sold on the market. Consequently, the construction of a replicating portfolio is impossible.

- The market is incomplete and illiquid. Most temperature-based markets are protected by limited trading activity and sparse price data. This lack of liquidity prevents the empirical calibration of risk-neutral measures based on observed market prices.

Due to these features, a commonly adopted approach in the academic and applied literature is to assume that the market price of the risk associated with the risk of temperature is zero, so the true measure coincides with the equivalent martingale measure. In this context, derivative prices are calculated as expected values based on the historical temperature distribution, often estimated via Monte Carlo simulation or historical burn analysis. This choice is motivated by the absence of reasons for the presence of a temperature risk premium in the literature. For example, in [1], it is chosen to use the zero-risk premium as well as in [2] and [8]. In such contexts the market price of risk, which we will henceforth call λ, is not directly observable and must be inferred from indirect sources. Several methods have been proposed in the literature to estimate this parameter, each based on different assumptions and available data. The most common approach reported in [11] is the calibration to market prices in which the value of λ is determined by fitting the market prices derived from the model to those observed. This inverse problem is typically solved by numerical optimization techniques but has the problem of requiring the availability of a sufficiently large and reliable dataset of derivative contracts. However, due to the limited liquidity of weather markets worldwide with the possible exception of the USA, this method is often unfeasible or produces unstable estimates. An example are the results obtained by Alaton in [1] in which using two contracts he obtains very different results. An alternative approach proposed in [5] involves the use of survey-based estimates of risk premiums. Some studies attempt to correlate the weather risk premium with macroeconomic factors or indicators of investor preferences, although this approach remains highly speculative and subject to significant model risk. For example, [2] justifies a hypothetical negative λ by telling us how investors agree to pay a premium as insurance to fix the variable temperature index rather than being exposed to its fluctuations. Because of the uncertainties associated with all these methods, many authors, particularly in the empirical literature, opt for the simplifying assumption of a zero risk premium ($\lambda = 0$), thus equating the price measure to the historical one.

4.2 Theoretical framework for pricing temperature derivatives

The valuation of weather derivatives relies on a probabilistic framework in which the payoff of a contract is modeled as a function of temperature indices, such as Heating Degree Days (HDD), Cooling Degree Days (CDD), or Cumulative Average Temperature (CAT). The definitions of these indices are provided in Chapter 1, in Definition 1.4.5 for HDD, Definition 1.4.3 for CDD, and Definition 1.4.7 for CAT. These indices are defined as path-dependent functionals of the temperature process T_t, and the payoff at maturity T is typically given by:

$$\text{payoff} = f(I_{[0,T]}),$$

where $I_{[0,T]}$ is a weather index accumulated over a contract period $[0, T]$.

We could define a relationship between the three indexes similar to the put-call parity.

Lemma 4.2.1.

$$CDD(\tau_1, \tau_2) \ - \ HDD(\tau_1, \tau_2) = CAT(\tau_1, \tau_2) \ - \ 18°C(\tau_2 - \tau_1).$$

Proof. Before proceeding we notice that

$$max(x - c, 0) - max(c - x, 0) = x - c.$$

- If $x \geq c$ then $max(x - c, 0) - max(c - x, 0) = x - c$;
- If $c \geq x$ then $max(x - c, 0) - max(c - x, 0) = -(c - x) = x - c$.

Let us now consider the left-hand side of the equality

$$\begin{aligned} CDD(\tau_1, \tau_2) \ - \ HDD(\tau_1, \tau_2) &= \sum_{i=\tau_1}^{\tau_2} max(T_{t_i} - 18°C, 0) - \sum_{i=\tau_1}^{\tau_2} max(18°C - T_{t_i}, 0) \\ &= \sum_{i=\tau_1}^{\tau_2} [max(T_{t_i} - 18°C, 0) - max(18°C - T_{t_i}, 0)] \\ &= \sum_{i=\tau_1}^{\tau_2} (T_{t_i} - 18°C) = \sum_{i=\tau_1}^{\tau_2} T_{t_i} - 18°C \cdot (\tau_1 - \tau_2) \\ &= CAT(\tau_1, \tau_2) \ - \ 18°C(\tau_2 - \tau_1). \end{aligned}$$

$\square$

The identity in Lemma 4.2.1 imposes a constraint that must also be reflected in arbitrage-free pricing models. Any pricing framework that simultaneously evaluates derivatives written on CDD, HDD, and CAT indices must preserve this relation under the risk-neutral expectation to ensure consistency and the absence of arbitrage.

The theoretical price V_t of a derivative with a temperature-based payoff under the risk-neutral measure $\mathbb{Q}$ is given by:

$$V_t = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[f(I_{[0,T]})|\mathcal{F}_t],$$

where r is the risk-free interest rate and the expectation is taken under the equivalent martingale measure and the filtration $\mathcal{F}_t$ is defined as in Equation (3.5).

Remark. Implicitly we assume that the HDDs, CDDs and the CATs are integrable respect to $\mathbb{Q}$. This assumption is justified by the physical nature of temperature, which ensures that the three indexes are bounded both below and above. Specifically, temperatures cannot fall below absolute zero (0° Kelvin) and are constrained by practical upper limits in real-world scenarios.

However, because temperature is a non-traded underlying, the market is incomplete and $\mathbb{Q}$ is not uniquely determined. In practice, especially in the absence of market-traded weather instruments, it is common to assume that the risk-neutral measure coincides with the physical measure $\mathbb{P}$, effectively assuming a zero market price of weather risk.

Thus, the pricing formula simplifies to: for $t < T$

$$V_t = e^{-r(T-t)}\mathbb{E}^{\mathbb{P}}[f(I_{[0,T]})|\mathcal{F}_t],$$

and it is typically computed via Monte Carlo simulation using temperature paths generated from a stochastic model (e.g., ARMA-GARCH or Ornstein-Uhlenbeck defined in Chapter 3) fitted on historical data.

4.3 Pricing by burn analysis

Burn analysis as reported in [3] is one of the widely used methods for evaluating and pricing weather derivatives. The popularity of this method is due to its practicality and simplicity, as the procedure consists of calculating the theoretical payoffs of a contract by applying the same conditions to the historical series. In practice, the pay-off is calculated for the previous years for which data is available. This process, called "burning" the contract on historical data, produces an empirical distribution of the pay-offs that represents an estimate of the risk and return profile of the contract itself. What makes it simple is the fact that it is a non-parametric method that does not make any type of preliminary assumptions on stochastic models or distributions for climate processes. This method is very useful in the case in which the climate indices are relatively stable and regular, such as temperatures. However, as seen, temperatures in recent years tend to increase, making this type of method not very suitable in our case. The method also has other limitations, especially when applied to other types of indices such as precipitation. The number of rainy days in a given period can change significantly from year to year, making the analysis less representative. Furthermore, as already announced, climate change and global meteorological phenomena can alter the future distribution of rainfall events, further reducing the reliability of models based only on the past. Finally, two other aspects to consider are the quantity and quality of the available data, since the effectiveness depends on the breadth of the historical series but also on the representativeness, as limited or too long series could compromise the accuracy of the estimate. Despite some limitations, burn analysis remains a fundamental tool for the evaluation of meteorological derivatives, as it can be considered as a starting point. To assess the predictive accuracy of the burn analysis method, we compare the 2024 observed values of HDDs from November to March and CDDs from May to September against their historical mean computed over the 2005–2023 period. In the Tables 4.1 and 4.2 we report for each city the observed 2024 value, the mean-based prediction, and the corresponding relative and percentage errors. This allows us to evaluate how closely the historical average approximates the actual 2024 outcome. The burn analysis procedure consists in collecting the historical values of the same type of contract over a selected past time window, in this case, the years 2005 to 2023, and computing their arithmetic mean. Specifically, for each city and for each index (HDD or CDD), we consider the values that would have been used to settle analogous contracts in the corresponding seasons of each previous year. These values are derived from the same daily temperature data series used for the forecast models. The arithmetic mean of these historical values is then taken as the prediction for the year 2024.

The Figures 4.1 and 4.2 illustrate the contract values for each year in the period

City	Observed 2024	Predicted (mean)	Percentage error (%)
Bologna	835.0	767.57	8.78
Bari	999.0	819.88	21.85
Bolzano	758.0	704.74	7.56
Viterbo	990.0	885.11	11.85
Roma	987.0	793.49	24.39
Palermo	1068.0	937.33	13.94
Ancona	851.0	671.88	26.66
Venezia	779.0	666.62	16.86
Torino	610.0	532.76	14.50

Table 4.1: Burn analysis results for CDDs May-September (2005-2023 vs 2024).

2005–2023, depicted as indigo stars. The observed contract values for the year 2024 are highlighted as pink points to clearly distinguish the most recent actual data. Additionally, the mean contract values obtained through the burn analysis method, calculated as the average of the historical values from 2005 to 2023, is shown as teal green stars and it represents the forecast for 2024. From the results and, in particular, from the Figures 4.1

City	Observed 2024	Predicted (mean)	Percentage error (%)
Bologna	1571.0	1742.73	9.85
Bari	974.0	1140.75	14.62
Bolzano	1799.0	1921.20	6.36
Viterbo	1130.0	1332.91	15.22
Roma	1026.0	1233.50	16.82
Palermo	482.0	608.24	20.75
Ancona	1269.0	1454.33	12.74
Venezia	1599.0	1729.25	7.53
Torino	1672.0	1879.23	11.03

Table 4.2: Burn analysis results for HDDs November-March (2005-2023 vs 2024).

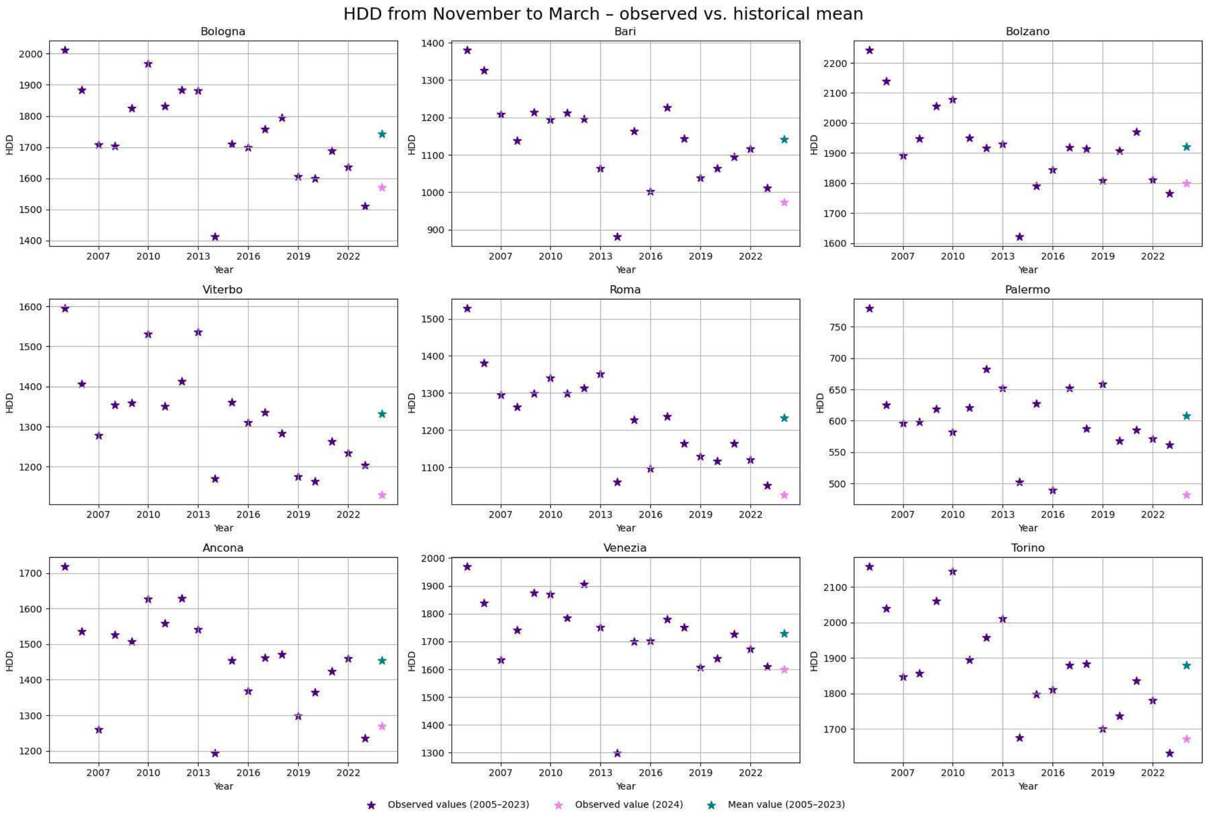


Figure 4.1: Time series of HDDs from November to March.

and 4.2 , it clearly emerges that the average of the last 19 years tends to underestimate the values of CDD and to overestimate those of HDD. This behavior is consistent with the expectations related to global warming and with the observations emerging from the analysis conducted in Chapter 2.

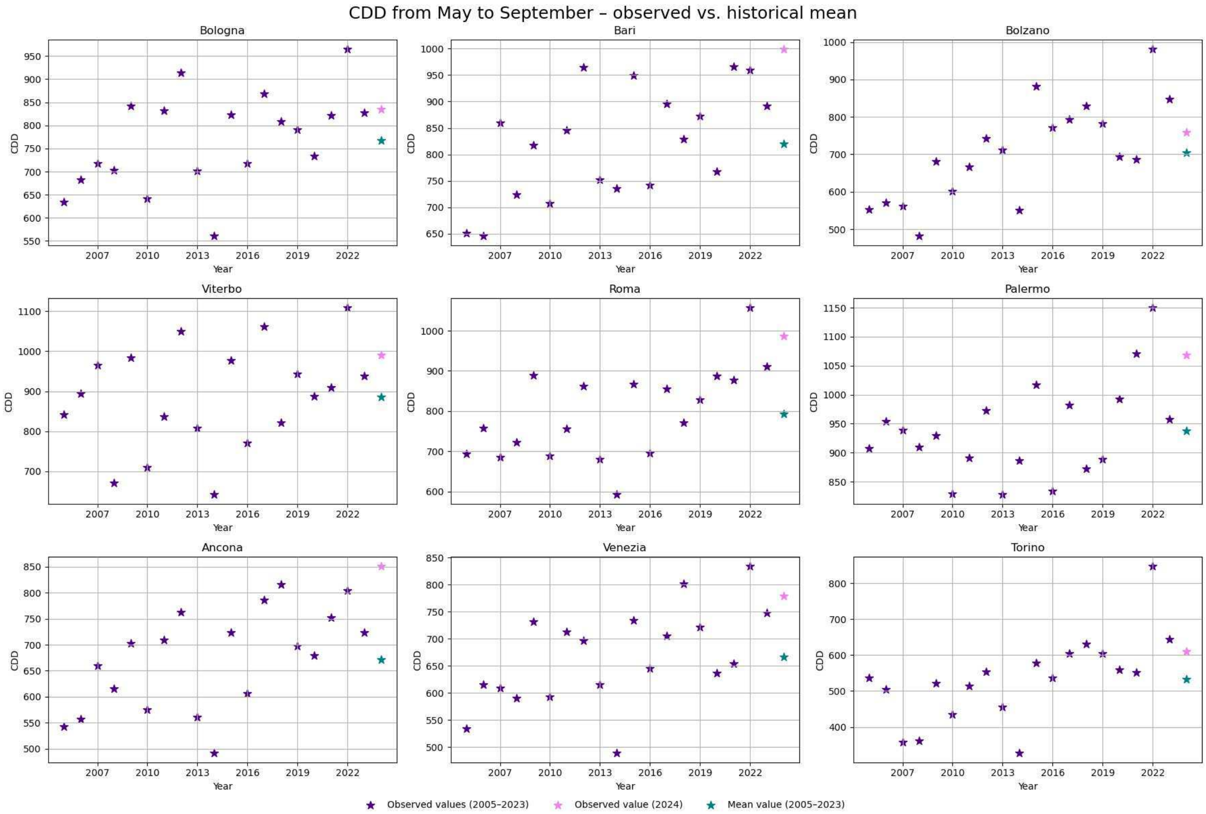


Figure 4.2: Time series of CDDs from May to September.

4.4 Derivative simulation under a modified Ornstein-Uhlenbeck temperature process

4.4.1 Approximate pricing of HDD options via max function simplification

Most weather derivatives involving temperature are based on Heating Degree Days or Cooling Degree Days, as introduced in Chapter 1. In this section, we present a standard approach to pricing an HDD call option using a simplification introduced by [1].

We consider an HDD call option with the payoff:

$$X = \max\{H_n - K, 0\}, \qquad \text{where} \quad H_n = \sum_{i=1}^{n} (18 - T_{t_i})^+.$$

The temperature process T_{t_i}, defined in Section 3.2.2, is modeled using a stochastic process which is the solution of the stochastic differential equation presented in Equation 3.6. As shown in Proposition 3.2.3, the solution to this process is normally distributed with a known mean and variance, which are derived in Proposition 3.2.2. We consider the pricing of a contract whose payout depends on the accumulation of Heating Degree Days over a specific winter period. As a concrete example, we focus on the month of January. In most Italian cities, average daily temperatures during January rarely exceed 18°C, which implies that the probability of the event $\max\{18 - T_{t_i}, 0\} = 0$ is very small. Based on this observation, we introduce a simplification that allows us to remove the maximum operator from the definition of HDD. Thus, we approximate the accumulated HDD over n days as:

$$H_n = \sum_{i=1}^{n}(18 - T_{t_i})^+ \approx 18n - \sum_{i=1}^{n} T_{t_i}.$$

To support this approximation, we report in Table 4.3 the empirical probabilities of daily average temperatures exceeding 18°C in January for Italian cities considered in this work. These probabilities are consistently low, justifying the removal of the maximum operator in the HDD calculation. To further assess the robustness of the HDD approximation,

City	Total Observations	Days Above 18°C	Probability
Bologna	620	0	0.00000
Bari	620	0	0.00000
Bolzano	620	0	0.00000
Viterbo	620	0	0.00000
Roma	620	0	0.00000
Palermo	620	7	0.01129
Ancona	620	0	0.00000
Venezia	620	0	0.00000
Torino	620	0	0.00000

Table 4.3: Empirical probability of daily average temperature exceeding 18°C in January, based on historical observations.

we report the empirical probabilities of daily average temperatures exceeding 18°C over extended winter periods. Table 4.4 presents the results for the five-month interval from November to March. While the approximation remains valid for most cities, a few exceptions emerge, particularly Palermo, and to a lesser extent Rome and Bari, where the number of days with temperatures above 18°C is no longer negligible. However, as shown in Table 4.5, if we restrict the analysis to the core winter months of December through February, the approximation holds very well for all cities except Palermo, where warm days are still observed with a non-trivial frequency.

Since T_{t_i}, for $i = 1, \dots, n$, are samples from an OU process taht is Gaussian process, the vector $(T_{t_1}, \dots, T_{t_n})$ is Gaussian. Since H_n is a linear combiantion of elements in the vector then it also follows a Gaussian distribution. The first and second order moments for $t < t_1$,

$$\mathbb{E}[H_n|\mathcal{F}_t] \approx \mathbb{E}\left[18n - \sum_{i=1}^{n} T_{t_i}\middle|\mathcal{F}_t\right] = 18n - \sum_{i=1}^{n} \mathbb{E}\left[T_{t_i}|\mathcal{F}_t\right] =: \mu_n \tag{4.1}$$

City	Total Observations	Days Above 18°C	Probability
Bologna	3025	2	0.00066
Bari	3025	30	0.00992
Bolzano	3025	3	0.00099
Viterbo	3025	5	0.00165
Roma	3025	16	0.00529
Palermo	3025	250	0.08265
Ancona	3025	10	0.00331
Venezia	3025	0	0.00000
Torino	3025	0	0.00000

Table 4.4: Empirical probability of daily average temperature exceeding 18°C during the November–March period.

City	Total observations	Days above 18°C	Probability
Bologna	1805	0	0.00000
Bari	1805	0	0.00000
Bolzano	1805	0	0.00000
Viterbo	1805	0	0.00000
Roma	1805	0	0.00000
Palermo	1805	40	0.02216
Ancona	1805	1	0.00055
Venezia	1805	0	0.00000
Torino	1805	0	0.00000

Table 4.5: Empirical probability of daily average temperature exceeding 18°C during the December–February period.

and

$$Var[H_n|\mathcal{F}_t] = \sum_{i=1}^{n} Var[T_{t_i}|\mathcal{F}_t] + 2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i>j}}^{n} Cov[T_{t_i}, T_{t_j}|\mathcal{F}_t] =: \sigma_n^2. \tag{4.2}$$

So, $H_n \sim N(\mu_n, \sigma_n^2)$ with μ_n the mean defined as in Equation (4.1) and the variance σ_n^2 defined in Equation (4.2).

Proposition 4.4.1. *The price at time $t \leq t_1$ for a HDD call option can be written as*

$$c(t) = e^{-r(t_n-t)}\left((\mu_n - K)\Phi\left(-\frac{K-\mu_n}{\sigma_n}\right) + \frac{\sigma_n}{\sqrt{2\pi}}e^{-\frac{(K-\mu_n)^2}{2\sigma_n^2}}\right).$$

Proof.

$$c(t) = e^{-r(t_n-t)}\mathbb{E}[(H_n-K)^+|\mathcal{F}_t] = e^{-r(t_n-t)}\int_K^{\infty}(x-K)\frac{1}{\sigma_n\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right)dx$$

Substituting

$$z = \frac{x-\mu_n}{\sigma_n}, \quad x = \mu_n + \sigma_n z, \quad dx = \sigma_n dz, \quad d = \frac{K-\mu_n}{\sigma_n}$$

we could rewrite

$$\begin{aligned}
c(t)e^{r(t_n-t)} &= \int_d^\infty (\mu_n + \sigma_n z - K)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz = \int_d^\infty (\mu_n - K)\phi(z)dz + \int_d^\infty \sigma_n z\phi(z)dz \\
&= (\mu_n - K)\int_d^\infty \phi(z)dz + \sigma_n \int_d^\infty z\phi(z)dz = (\mu_n - K)[1-\Phi(d)] + \sigma_n\phi(d) \\
&= (\mu_n - K)\Phi(-d) + \sigma_n\phi(d) = (\mu_n - K)\Phi\left(\frac{\mu_n - K}{\sigma_n}\right) + \sigma_n\phi\left(\frac{K-\mu_n}{\sigma_n}\right) \\
&= (\mu_n - K)\Phi\left(-\frac{K-\mu_n}{\sigma_n}\right) + \frac{\sigma_n}{\sqrt{2\pi}}e^{-\frac{(K-\mu_n)^2}{2\sigma_n^2}}
\end{aligned}$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ is the standard normal density function and $\Phi(z) = \int_{-\infty}^z \phi(t)dt$ is the cumulative distribution function. □

Proposition 4.4.2. *The price at time $t \leq t_1$ for a HDD put option can be written as follows*

$$p(t) = e^{-r(t_n-t)}\left[(K-\mu_n)\left(\Phi(d) - \Phi\left(-\frac{\mu_n}{\sigma_n}\right)\right) + \frac{\sigma_n}{\sqrt{2\pi}}\left(e^{-\frac{d^2}{2}} - e^{-\frac{1}{2}\left(\frac{\mu_n}{\sigma_n}\right)^2}\right)\right]$$

Proof.

$$p(t) = e^{-r(t_n-t)}\mathbb{E}[(K-H_n)^+ \mid \mathcal{F}_t] = e^{-r(t_n-t)}\int_{-\infty}^K (K-x)\frac{1}{\sigma_n\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right)dx$$

Substituting

$$z = \frac{x-\mu_n}{\sigma_n}, \quad x = \mu_n + \sigma_n z, \quad dx = \sigma_n dz, \quad d = \frac{K-\mu_n}{\sigma_n}$$

we rewrite

$$\begin{aligned}
p(t)e^{r(t_n-t)} &= \int_0^d (K - \mu_n - \sigma_n z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz = (K-\mu_n)\int_0^d \phi(z)dz - \sigma_n\int_0^d z\phi(z)dz \\
&= (K-\mu_n)\left(\Phi(d) - \Phi\left(-\frac{\mu_n}{\sigma_n}\right)\right) - \sigma_n\left[-\phi(d) + \phi\left(-\frac{\mu_n}{\sigma_n}\right)\right] \\
&= (K-\mu_n)\left(\Phi(d) - \Phi\left(-\frac{\mu_n}{\sigma_n}\right)\right) + \frac{\sigma_n}{\sqrt{2\pi}}\left(e^{-\frac{d^2}{2}} - e^{-\frac{1}{2}\left(\frac{\mu_n}{\sigma_n}\right)^2}\right)
\end{aligned}$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ is the standard normal density function and $\Phi(z) = \int_{-\infty}^z \phi(t)dt$ is the cumulative distribution function. □

The same approximation cannot be applied to options on Cooling Degree Days, since the probability that the daily average temperature falls below 18°C during the summer months is not negligible, as illustrated in Table 4.6. In this case, the simplification

$$C_n = \sum_{i=1}^{n} (T_{t_i} - 18)^+ \approx \sum_{i=1}^{n} T_{t_i} - 18n$$

would not be reliable, because the probability that $(T_{t_i} - 18)^+ = 0$ is not sufficiently small. Therefore, the contribution of days with temperatures below the threshold cannot be ignored in the calculation.

City	Total observations	Days below 18°C	Probability
Bologna	1840	23	0.0125
Bari	1840	22	0.01196
Bolzano	1840	50	0.02717
Viterbo	1840	26	0.01413
Roma	1840	21	0.01141
Palermo	1840	3	0.00163
Ancona	1840	24	0.01304
Venezia	1840	25	0.01359
Torino	1840	70	0.03804

Table 4.6: Number of days with temperature below 18°C from June to August.

4.4.2 Monte Carlo simulation for HDD and CDD

In the case of options on climate indices based on temperature, such as HDD and CDD, it is not always possible to derive analytical formulas for pricing, especially in the case of a realistic and complex model to describe temperatures. In cases like this as suggested in [10], the choice very often falls on the Monte Carlo method which involves the use of numerous simulations to approximate the expected value of the option payoff even in the presence of complex dynamics such as those analyzed in this paragraph. The fundamental step to price an option is to find the conditional expected value under the equivalent martingale measure of the HDD or CDD. In this particular case the equivalent martingale measure coincide with the real world probability.

Definition 4.4.1 (Monte Carlo pricing of weather derivatives). *Let $f(X)$ denote the payoff function of a weather derivative written on a temperature index X (e.g., HDD, CDD or CAT) computed over a time interval $[\tau_1, \tau_2]$. Assuming a risk-free interest rate r and working under the historical probability measure $\mathbb{P}$ (in the absence of a liquid market for weather derivatives), the fair price of the derivative at time $t \leq \tau_1$ is given by:*

$$V_t = \mathbb{E}^{\mathbb{P}} \left[e^{-r(\tau_2 - t)} f(X(\tau_1, \tau_2)) \,\middle|\, \mathcal{F}_t \right],$$

where $\mathcal{F}_t$ denotes the available information at time t.

Since the distribution of X is generally unknown in closed form, the expectation is approximated via Monte Carlo simulation. Given M independent simulations $\{X^{(m)}\}_{m=1}^{M}$ of the index X, the price can be estimated as:

$$V_t \approx \frac{e^{-r(\tau_2-t)}}{M} \sum_{m=1}^{M} f\left(X^{(m)}\right). \tag{4.3}$$

To find it, the following procedure was followed:
For each simulation:

1. For each Monte Carlo iteration, a complete trajectory of daily temperatures is simulated over the entire duration of the option,
2. based on the simulated path, the cumulative value of the relevant index (such as CAT, HDD, or CDD) is computed,
3. this procedure is repeated for a large number of independent simulations (20000 in our case) to approximate the distribution of the payoff.

Finally, the option price is estimated as the arithmetic mean of the payoffs obtained as in Equation (4.3).

To simulate the temperature trajectory, the explicit analytical solution of the process is used. The process is simulated by iterating along a discrete time interval, taking one day at a time, starting from a known initial value, the temperature of the last observed day, and calculating the simulated temperature for each subsequent day. The simulation horizon depends on the usage, for example for an HDD option referring to the period November 2023 – March 2024, the simulated temperatures cover exactly that interval and use as data to train the model those up to the previous day. In this way the simulations are completely out-of-sample. The simulations for the first model focus exclusively on the values of the HDD or CDD indices rather than on the option prices themselves. This is justified by the fact that, under the assumption of constant interest rate, the pricing involves only a deterministic discounting operation. Furthermore, in the case of standard options such as calls, the strike price K is fixed and does not introduce additional uncertainty. As a result, the stochastic component of the option price is entirely driven by the behavior of the index. Therefore, validating the model with respect to the index values is sufficient to assess its predictive accuracy, as a good performance on the index naturally translates into reliable option pricing. To illustrate the convergence behavior of the Monte Carlo estimator, Figure 4.3 compares the estimated price of a representative HDD contract using 20, 200, 2000, and 20000 simulations. The complete code for the simulation which results are reported in the Tables 4.7 and 4.8, the training of the OU model and the calculation of the HDD/CDD indices is reported in the Appendix. Regarding computational efficiency,

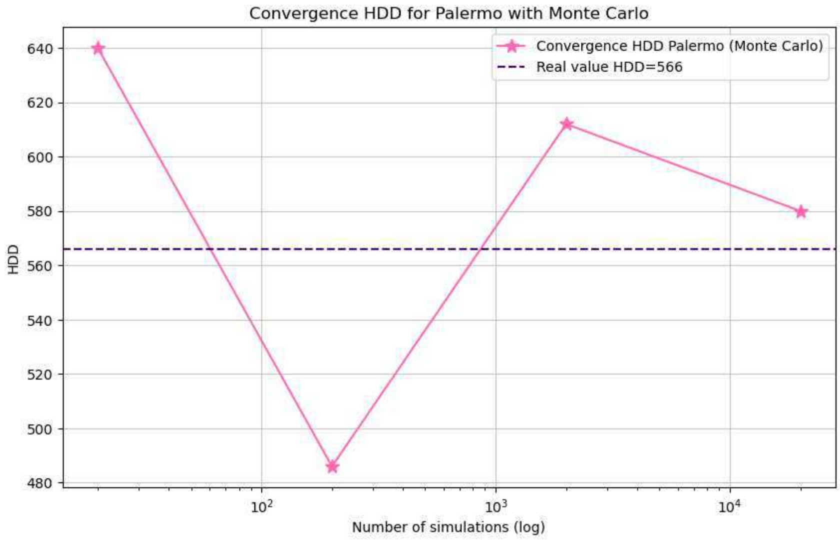


Figure 4.3: Convergence of Monte Carlo estimates for HDD pricing under the OU model with increasing number of simulations for the city of Palermo.

the simulation of 20000 paths for a single city, using already estimated model parameters, required approximately 3 minutes on a standard desktop machine. This demonstrates the feasibility of large-scale simulation studies for multiple locations without excessive computational cost.

City	Observed CDD	Simulated CDD	Error	Percentage error (%)
Torino	610	655	45	7.38
Roma	987	913	-74	-7.50
Bologna	835	874	39	4.67
Viterbo	990	940	-50	-5.05
Palermo	1068	984	-84	-7.87
Bari	999	910	-89	-8.91
Venezia	779	738	-41	-5.26
Ancona	851	755	-96	-11.29
Bolzano	758	848	90	11.87

Table 4.7: Comparison between observed and simulated CDD for the period May-September 2024, with absolute and percentage error - Modified OU process.

4.5 Derivative simulation under a ARMA-GARCH temperature process

4.5.1 Monte Carlo simulation for HDD and CDD

The pricing procedure adopted under the ARMA–GARCH framework follows the same Monte Carlo methodology previously described for the Ornstein–Uhlenbeck process. To

City	Observed HDD	Simulated HDD	Error	Percentage error (%)
Bari	902	1064	162	17.96
Bologna	1473	1649	176	11.95
Palermo	566	580	14	2.47
Viterbo	1055	1068	13	1.23
Venezia	1567	1658	91	5.81
Torino	1618	1733	115	7.11
Bolzano	1767	1766	-1	-0.06
Ancona	1166	1064	-102	-8.75
Roma	916	1065	149	16.26

Table 4.8: Comparison of observed and simulated HDDs for the period November-March 2024, with absolute and percentage errors- Modified OU process.

avoid redundancy, we refer the reader to Section 4.4.2 for the detailed exposition of the simulation-based pricing scheme. In this case, the temperature paths used for pricing were generated using the ARMA–GARCH model introduced in Sections 3.3.2 and 3.3.3, with parameters fixed to those previously estimated on historical data. This choice, while simplifying implementation and reducing computational overhead, preserves the statistical characteristics of the model calibrated in the initial step. Due to the presence of time-varying volatility and the recursive nature of the GARCH component, the simulation of temperature trajectories under the ARMA–GARCH model is substantially more computationally demanding. In practice, the average execution time for the pricing simulations was found to be approximately ten times higher or more than that required by the OU-based approach, for the same number of Monte Carlo replications. Nevertheless, the valuation remains coherent with the general pricing framework adopted in this study. As with the OU model, pricing was performed under the historical probability measure, assuming a zero market price of risk. To illustrate the convergence behavior of the Monte Carlo estimator under the ARMA–GARCH dynamics, Figure 4.4 compares the estimated price of a representative CDD contract using 20, 200, 2000, and 20000 simulations. As expected, the variance of the estimator diminishes with increasing sample size, and convergence is reached around 20000 simulations.

The results obtained for the Cooling Degree Days and Heating Degree Days are reported in the Tables 4.9 and 4.10. As the discounting factor plays only a deterministic role under the assumption of a constant interest rate, and since the structure of a vanilla option is typically deterministic, the simulations focus exclusively on the stochastic behavior of the indices rather than on option pricing, as already done for the first model. Moreover, given the absence of actual market prices for the options under consideration, any comparison would be of limited value, unlike in the case of temperature accumulations, for which real data is available. The full implementation details of the Monte Carlo pricing routine, including the generation of ARMA–GARCH trajectories, are provided in

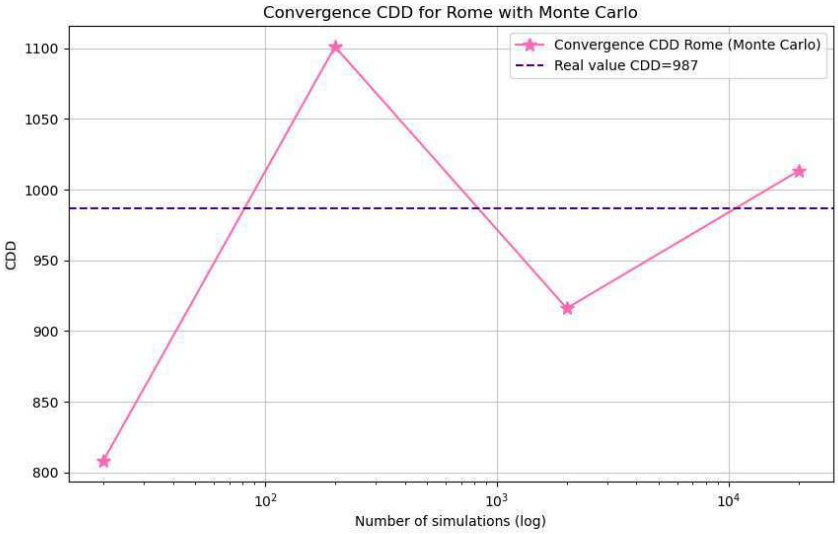


Figure 4.4: Convergence of Monte Carlo estimates for CDD pricing under the ARMA–GARCH model with increasing number of simulations for the city of Rome.

Appendix A.9.

City	Observed HDD	Simulated HDD	Error	Percentage error (%)
Bari	902	911	9	1.00
Bologna	1473	1305	-168	-11.41
Palermo	466	453	-13	-2.79
Viterbo	1055	1005	-50	-4.74
Venezia	1567	1342	-225	-14.36
Torino	1618	1606	-12	-0.74
Bolzano	1768	1829	61	3.45
Ancona	1166	1036	-130	-11.15
Roma	916	897	-19	-2.07

Table 4.9: Comparison between observed and simulated CDD for the period May-September 2024, with absolute and percentage error - ARMA-GARCH process.

City	Observed CDD	Simulated CDD	Error	Percentage error (%)
Bari	999	924	-75	-7.51
Palermo	1068	978	-90	-8.42
Torino	610	593	-17	-2.79
Roma	987	1013	26	2.63
Bologna	835	899	64	7.66
Bolzano	758	852	94	12.40
Viterbo	990	970	-20	-2.02
Venezia	779	767	-12	-1.54
Ancona	851	787	-64	-7.52

Table 4.10: Comparison between observed and simulated CDD for the period May-September 2024, with absolute and percentage error - ARMA-GARCH process.

4.6 Comments on the results

From the two Figures 4.5 and 4.6 we can say that, in general, the ARMA - GARCH model proves superior in precision for both CDD and HDD index simulations. Its superiority can be justified in its ability to capture dynamic volatility since in the OU model the volatility itself is simplified to constant values over a month. The OU model, although simpler and potentially robust in some contexts, shows lower precision, especially when the time series present greater complexity or seasonal variability. The Burn analysis, as expected, is the least precise in both indices. However, its use was intended to validate the models, as it represents an easy and inexpensive method at a computational level. Its imprecision is due to the lower local specificity and in part to ongoing climate changes since it is not able to capture the evolution of temperatures. We conclude that the evident variation in model performance varies significantly from city to city, underlining the importance of calibrating and testing models on specific local data, since climate conditions and their dynamics can differ significantly even within the same country.

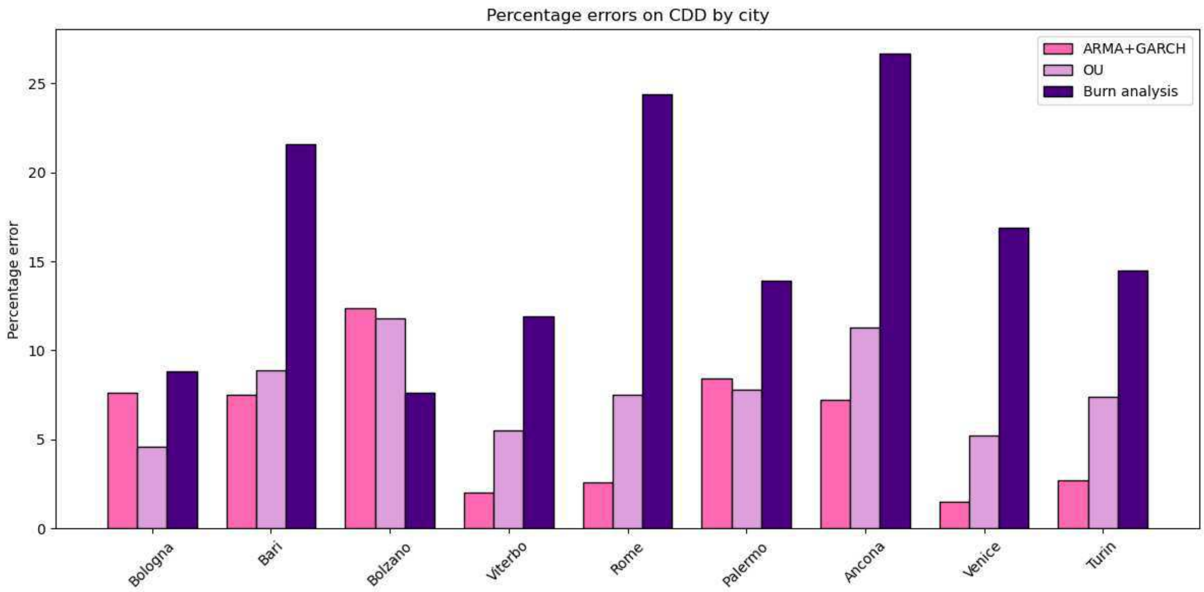


Figure 4.5: Comparison of CDD prediction errors: ARMA+GARCH, OU and Burn Analysis.

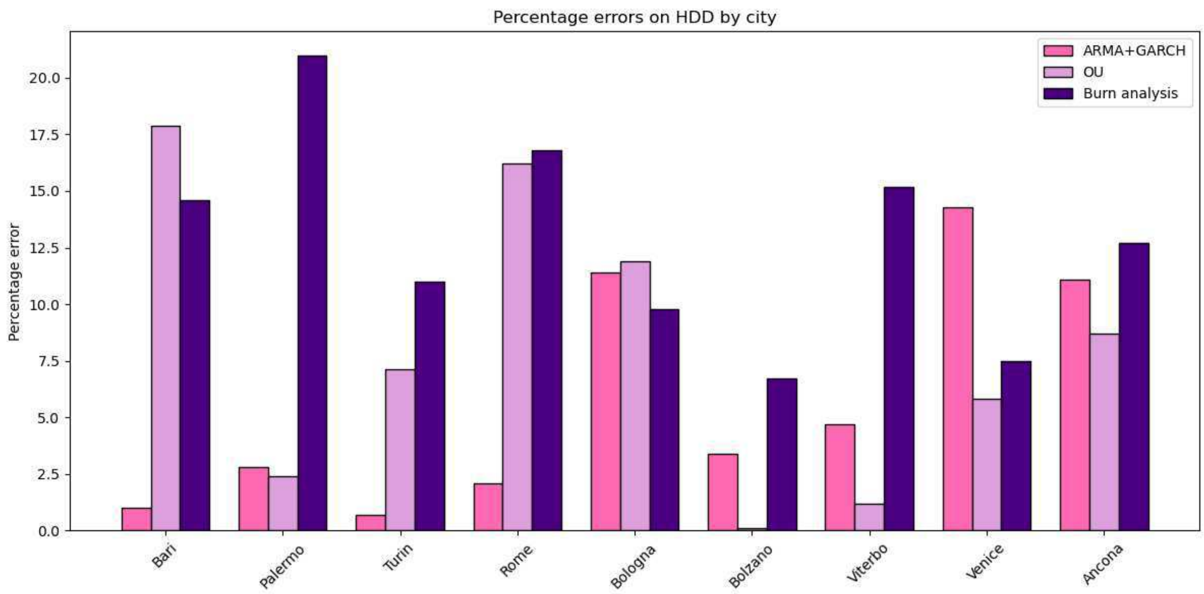


Figure 4.6: Comparison of HDD prediction errors: ARMA+GARCH, OU and Burn Analysis.

Conclusion

This thesis has investigated the use of weather derivatives in the Italian context, with a particular focus on temperature-based contracts. The preliminary analysis of historical temperature data from nine cities, selected to represent Italy's diverse climate zones, provided the necessary foundation to develop realistic and robust forecasting models. This exploratory phase was essential to ensure that the modeling assumptions aligned with the actual behavior of the temperature data across the country.

The core of the work was the stochastic modeling of temperatures. Two different models were developed and then compared. The first model represented by a continuous-time stochastic process is a generalization of the Ornstein-Uhlenbeck to which the constant mean was replaced with a time-dependent deterministic function. The second model, an ARMA–GARCH, instead worked on the residuals of the series after removing the seasonal component. Both models were calibrated using series of daily mean temperature from 2005 to 2023. The choice to use data up to 2023 is due to the use of 2024 to test the predictive capacity of the model. The results obtained in terms of performance are very similar between the two models. It should also be taken into account that the model using the modified OU process simplifies the value of volatility to a constant monthly value, unlike the second model which, thanks to the GARCH component, also provides a dynamic to volatility. Despite this limitation of the first model, the results do not differ greatly, making the monthly approximation a good choice, also due to the lower computational cost. The reason for the accurate statistical modeling of daily mean temperatures lies in the objective set at the beginning of the investigation of the use of temperature-based weather derivative contracts. These contracts are based on indices that summarize the thermal trend of a given period, such as Heating Degree Days (HDD) and Cooling Degree Days (CDD), whose value depends directly on the observed temperatures. A robust and consistent modeling of temperatures is therefore essential to be able to reliably estimate these indices and, consequently, for a correct evaluation and pricing of the related derivative instruments. Once the models were developed, a simulation phase was conducted with the aim of predicting the trend of the HDD and CDD indices for the year 2024. Using both models, synthetic trajectories of the daily mean temperatures were generated from which the numerical values of the indices were then obtained. The values obtained were subsequently compared with those actually observed during the year 2024.

The burn analysis method was used as a term of comparison. The comparison clearly highlighted the better performance of the proposed stochastic models compared to the chosen benchmark. Between the two models, the one based on the modified Ornstein-Uhlenbeck process has a clear advantage in computational terms: it is able to complete 20,000 simulations in about 3 minutes for each city, compared to the approximately 30 minutes required by the ARMA-GARCH model. While the latter generally offers slightly superior performance in terms of predictive accuracy, the OU model proves particularly useful in contexts where a trade-off between accuracy and speed of execution is necessary. The work has certain limitations. First, the models used are purely statistical and do not incorporate weather forecast data or exogenous variables, which could improve predictive accuracy. Second, the lack of a liquid and transparent weather derivatives market in Italy prevents the direct estimation of the market price of risk, which limits the practical applicability of pricing results.

Nevertheless, this thesis offers a solid framework for the analysis and modeling of temperature in the Italian context and its application to financial instruments. Future research could explore hybrid models that incorporate meteorological forecasts or machine learning techniques, as well as extend the study to other weather variables in the Italian context such as precipitation or wind speed. Furthermore, increased market development in this sector would provide an opportunity for further empirical validation of the pricing models presented.

Chapter A

Code snippets

```python
import os
import pandas as pd

directory = 'C:\\Users\\arida\\Desktop\\tesi\\dati\\venezia\\'

#function to extract year and month from file name
def estrai_data(nome_file):
    parti = nome_file.split('-')
    anno = int(parti[1])
    mese = parti[2].replace('.csv', '')
    mesi = { 'Gennaio': 1, 'Febbraio': 2, 'Marzo': 3, 'Aprile': 4, '
        Maggio': 5, 'Giugno': 6, 'Luglio': 7, 'Agosto': 8, 'Settembre':
        9, 'Ottobre': 10, 'Novembre': 11, 'Dicembre': 12 }
    return anno, mesi[mese]

dataframes = []
#scan of the files in the directory
for file in os.listdir(directory):
    if file.startswith('Venezia') and file.endswith('.csv'):
        # Estrai la data dal nome del file
        anno, mese = estrai_data(file)
        df = pd.read_csv(os.path.join(directory, file), sep=';')
        df['Anno'] = anno
        df['Mese'] = mese
        dataframes.append(df)

df_completo = pd.concat(dataframes)
#order the data using the date
df_completo.sort_values(by=['Anno', 'Mese'], inplace=True)
#save the new dataframe in a excel file
df_completo.to_excel('venezia.xlsx', index=False)
```

Listing A.1: Code to merge all files for the city Venice

```
#mean of 7 days before and after for Tmin and Tmax
rolling_min = data['TMIN C'].shift(-7).rolling(window=14,min_periods=1).
    mean()
rolling_max = data['TMAX C'].shift(-7).rolling(window=14,min_periods=1).
    mean()

#NaN values replaced with the local mean for Tmin and Tmax
data['TMIN C'] = data['TMIN C'].fillna(rolling_min)
data['TMAX C'] = data['TMAX C'].fillna(rolling_max)

#replace NaN values in the mean temperature with the mean of Tmax and
    Tmin for that day
data['TMEDIA C'] = data.apply(lambda row: (row['TMAX C'] + row['TMIN C'
    ]) / 2 if pd.isna(row['TMEDIA C']) else row['TMEDIA C'], axis=1)
```

Listing A.2: Code to handle NaN values

```
colonne=['TMEDIA C','TMIN C','TMAX C','PUNTORUGIADA C','UMIDITA %','
    VISIBILITA km','VENTOMEDIA km/h','VENTOMAX km/h','RAFFICA km/h', '
    PRESSIONESLM mb', 'PRESSIONEMEDIA mb']
for c in colonne:
    data[c] = data[c].astype(str)

#clean the data
    data[c] = data[c].str.strip()  # remove spaces in the string
    data[c]=data[c].str.replace(r'[^\d.,-]', '', regex=True)  # remove
        everything that can't be used to define a number
    data[c] = data[c].str.replace(',', '.')  # substitute the comma with
         the full stop

#convert to float
    data[c] = pd.to_numeric(data[c], errors='coerce')
```

Listing A.3: Code to clean the data

```
from pykrige.ok import OrdinaryKriging
city_coords = {
    "viterbo": (42.4207, 12.1077),
    "roma": (41.9028, 12.4964),
    "ancona": (43.6158, 13.5189),
    "venezia": (45.4408, 12.3155),
    "bari": (41.1171, 16.8719),
    "bolzano": (46.4983, 11.3548),
    "torino": (45.0703, 7.6869),
    "bologna": (44.4949, 11.3426),
    "palermo": (38.1157, 13.3615)
}
```

```
cities = ["viterbo", "roma", "ancona", "venezia", "bari", "bolzano", "
    torino", "bologna", "palermo"]

temps = []
lats = []
lons = []

for city in cities:
    df = globals()[city]
    temp = df[df['mese'].isin([12,1,2])]["TMEDIA C"].mean()
    lat, lon = city_coords[city],,

    temps.append(temp)
    lats.append(lat)
    lons.append(lon)

OK = OrdinaryKriging(lons, lats, temps, variogram_model='spherical')

grid_lon = np.linspace(6, 19, 100)
grid_lat = np.linspace(36, 47.5, 100)
gridx, gridy = np.meshgrid(grid_lon, grid_lat)

z, ss = OK.execute('grid', grid_lon, grid_lat)

z, ss = OK.execute('grid', grid_lon, grid_lat)
plt.figure(figsize=(10, 6))
plt.contourf(gridx, gridy, z, cmap='RdPu', levels=15)
plt.scatter(lons, lats, c=temps, cmap='RdPu', edgecolor='black', s=100)
for i, city in enumerate(cities):
    city_name = city.capitalize()  # Iniziale maiuscola
    plt.text(lons[i]+0.1, lats[i], city_name, fontsize=9)
plt.colorbar(label="Mean temperature")
plt.xlim(6, 19)
plt.ylim(36, 47.5)
plt.title("Interpolated Heat Map: Winter")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.grid(True)
plt.show()
```

Listing A.4: Code to perform the Kriging

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
import scipy.stats as sp
```

```python
from scipy.stats import jarque_bera
from scipy.stats import norm

def modello(citta, nome):
    y = np.array(citta['TMEDIA C'])
        t = np.arange(len(y)) + 1
        w = 2 * np.pi / 365
        X = pd.DataFrame({
                'const': 1,
                't': t,
                'sin_wt': np.sin(w * t),
                'cos_wt': np.cos(w * t)
        })
        model = sm.OLS(y, X).fit()
        print(model.summary())
        residui = model.resid

        fig, axes = plt.subplots(1, 2, figsize=(15, 5))
        # histogram with density curve
        count, bins, _ = axes[0].hist(residui, bins=30, density=True, alpha=0.6, color='violet', edgecolor='black', label="Resid")
        sns.kdeplot(residui, color="violet", lw=2, label="Log returns density estimation", ax=axes[0])
        x = np.linspace(bins[0], bins[-1], 1000)
        y = norm.pdf(x, residui.mean(), residui.std())
        axes[0].plot(x, y, color='indigo', lw=2, label='Normal density')
        axes[0].set_title(f'Histogram with density curve for residuals of {nome} model')
        axes[0].set_xlabel('Log Returns')
        axes[0].set_ylabel('Density')
        axes[0].legend()

        #QQ-Plot with normal distribution
        sp.probplot(residui, dist="norm", plot=axes[1])  # Correct usage of probplot
        axes[1].set_title("QQ-Plot for normality")
        axes[1].get_lines()[0].set_color('violet')
        axes[1].get_lines()[1].set_color('indigo')
        axes[1].grid(True)

        plt.tight_layout()
        plt.show()

        #JB test
        jb_stat, p_value = jarque_bera(residui)
        if p_value < 0.05:
        print("The residuals are NOT normally distributed.")
```

```
        else:
        print("The residuals can be considered normal.")

        return model

def calcolo_parametri(df):
        A=[]
        B=[]
        C=[]
        phi=[]
        for c in df.columns:
        serie=df[c]
        A.append(serie['const'])
        B.append(serie['t'])
        C.append(np.sqrt(serie['sin_wt']**2+serie['cos_wt']**2))
        phi.append(np.arctan(serie['cos_wt']/serie['sin_wt'])-np.pi)
        return A,B,C,phi
        parametri=calcolo_parametri(df)
        params=pd.DataFrame(parametri)
        params.index = ['A', 'B', 'C','phi']
        params.columns=nomi
        def model(x, parametri):
        A=parametri['A']
        B=parametri['B']
        C=parametri['C']
        phi=parametri['phi']
        return A + B * x + C * np.sin(2 * np.pi * x / 365 + phi)

def model_plot(nome,citta):
        parti = partial(model, parametri=params[nome])
        x_values = np.arange(1, len(citta[citta['anno'] < 2024]['TMEDIA
            C']) + 1)
        y_values = list(map(parti, x_values))
        fig, axes = plt.subplots(1, 2, figsize=(14, 5), sharey=True)

        #first plot
        axes[0].plot(citta['DATA'], citta['TMEDIA C'], color='plum',
            label='Observed')
        axes[0].plot(citta['DATA'], y_values, color='indigo', label='
            Model')
        axes[0].set_title(f'Temperatures vs Model for {nome}')
        axes[0].set_xlabel('Date')
        axes[0].set_ylabel('Average Temperature')
        axes[0].legend()

        #second plot: zoom 2005
        mask_2005 = citta['anno'] == 2005
```

```
        citta_2005 = citta[mask_2005]
        axes[1].plot(citta_2005['DATA'], citta_2005['TMEDIA C'], color='plum', label='Observed')
        axes[1].plot(citta_2005['DATA'], y_values[:365], color='indigo', label='Model')
        axes[1].set_title(f'Temperatures vs Model for {nome} - Zoom 2005')
        axes[1].set_xlabel('Date')
        axes[1].legend()

        plt.tight_layout()
        plt.show()

        return y_values
```

Listing A.5: Code to calculate the model parameters

```
        import matplotlib.pyplot as plt
        from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
        from statsmodels.tsa.stattools import adfuller


cols = diff_df.columns
n = len(cols)
rows, cols_per_row = 3, 3
global_min = diff_df.min().min()-2
global_max = diff_df.max().max()+2

fig, axes = plt.subplots(rows, cols_per_row, figsize=(15, 10))
axes = axes.flatten()
for i, col in enumerate(diff_df.columns):
        ax = axes[i]
        ax.plot(diff_df[col], color='hotpink')
        ax.set_title(f'Stochastic part - {col}')
        ax.set_xlabel('Date')
        ax.set_ylim(global_min, global_max)

for j in range(i+1, len(axes)):
        fig.delaxes(axes[j])
        plt.tight_layout()
        plt.show()

def adf_test_all_columns(df, signif=0.05):
        results = []

        print(f"{'Colonna':<20} {'ADF Statistic':>15} {'p-value':>10} {'Conclusione':>20}")
        print("-" * 70)
```

```

        for col in df.columns:
        series = df[col].dropna()
        result = adfuller(series)
        adf_stat = result[0]
        p_value = result[1]

        if p_value < signif:
        conclusion = "Stazionaria"
        else:
        conclusion = "Non stazionaria"

        print(f"{col:<20} {adf_stat:>15.4f} {p_value:>10.4f} {conclusion:>20}")
        results.append({
                'colonna': col,
                'adf_stat': adf_stat,
                'p_value': p_value,
                'conclusione': conclusion
        })

        return results


def plot_acf_pacf(series, lags=20, title=''):
        fig, axes = plt.subplots(1, 2, figsize=(12, 4))

        plot_acf(series.dropna(), lags=lags, ax=axes[0], color='blueviolet')
        axes[0].set_title(f'ACF - {title}')

        plot_pacf(series.dropna(), lags=lags, ax=axes[1], method='ywm', color='blueviolet')
        axes[1].set_title(f'PACF - {title}')

        plt.tight_layout()
        plt.savefig(f'{title}_ACF')
        plt.show()
for c in diff_df.columns:
        plot_acf_pacf(diff_df[c], lags=20, title=c)
```

Listing A.6: Code to investigate the residuals before the ARMA model definition

```
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

```

```
#orders from auto_arima
orders = {
                'Bari': (2, 2),
                'Palermo': (2, 3),
                'Bologna': (3, 1),
                'Ancona': (3, 1),
                'Torino': (2, 3),
                'Roma': (2, 2),
                'Viterbo': (1, 0),
                'Venezia': (1, 1),
                'Bolzano': (2, 1)
        }

params_df = pd.DataFrame()

for city, (p, q) in orders.items():
        series = diff_df[city]
        try:

                model = ARIMA(series, order=(p, 0, q)).fit()
                params = model.params.to_dict()
                city_params_df = pd.DataFrame(params, index=[city])
                params_df = pd.concat([params_df, city_params_df], axis
                   =0)
                residuals = model.resid

                #ACF e PACF dei residui
                fig, axes = plt.subplots(1, 2, figsize=(12, 4))
                plot_acf(residuals, lags=20, ax=axes[0])
                axes[0].set_title(f'ACF dei residui - {city}')
                plot_pacf(residuals, lags=20, ax=axes[1], method='ywm')
                axes[1].set_title(f'PACF dei residui - {city}')
                plt.tight_layout()
                plt.savefig(f'residui_acf_pacf_{city}.png')
                plt.close()

        except Exception as e:
            print(f"Errore per {city}: {e}")

print(params_df)
```

```
#deterministic part
def simulate_temperature_trajectory(start_date, end_date, A, B, C, omega
   , phi, x0):
        dates = pd.date_range(start=start_date, end=end_date, freq='D')
        N = len(dates)
        dt = 1.0
```

```
        t_vals = np.arange(6939, 7305)
        Tm_t = A + B * t_vals + C * np.sin(omega * t_vals + phi)
        return pd.Series(data=Tm_t, index=dates)
params=pd.read_excel('19.xlsx')
params.index=params['Unnamed: 0']
params = params.drop('Unnamed: 0', axis=1)
deterministic=pd.DataFrame()
citta=[bologna,bari,bolzano,viterbo,roma,palermo,ancona,venezia,torino]
for i,c in enumerate(params.columns):
        c2=citta[i]
        A = params[c]['A']
        B = params[c]['B']
        C = params[c]['C']
        omega = 2 * np.pi / 365
        phi = params[c]['phi']
        x0 = float(c2[c2.index=='2023-12-31']['TMEDIA C'] )
        Tm=simulate_temperature_trajectory('2024-01-01', '2024-12-31', A
            , B, C, omega, phi, x0)
        deterministic[c]=Tm
```

```
from sklearn.metrics import mean_absolute_error, mean_squared_error
from scipy.stats import pearsonr

def forecast(p,q,name,df,pl=0):
        det=deterministic[name]
        series=diff_df[name]
        arma_model = ARIMA(series, order=(p, 0, q))
        arma_fit = arma_model.fit()
        arma_resid = arma_fit.resid
        garch_model = arch_model(arma_resid, vol='Garch', p=1, q=1, dist
            ='t')
        garch_fit = garch_model.fit(disp="off")

        #dynamic forecast
        def forecast_arma_garch_iterative(arma_fit, garch_fit, steps):
                forecast_values = []
                volatility_values = []
                last_observed_value = arma_fit.fittedvalues[-1]
                last_volatility = garch_fit.conditional_volatility[-1]

                for step in range(steps):
                forecast_step = arma_fit.forecast(steps=1)[0]
                volatility_step = garch_fit.forecast(horizon=1).variance
                    .values[-1, 0]
                nu = garch_fit.params['nu']
                forecast_with_volatility = forecast_step + np.random.
                    standard_t(df=nu) * np.sqrt(volatility_step)
```

```
                forecast_values.append(forecast_with_volatility)
                volatility_values.append(np.sqrt(volatility_step))
                last_observed_value = forecast_with_volatility

                return np.array(forecast_values), np.array(volatility_values)

        forecast_values, volatility_values = forecast_arma_garch_iterative(arma_fit, garch_fit, 365)
        forecast_index = pd.date_range(start=series.index[-1], periods=366, freq='D')[1:]
        forecast_df = pd.DataFrame(forecast_values, index=forecast_index, columns=['Forecast'])
        volatility_df = pd.DataFrame(volatility_values, index=forecast_index, columns=['Volatility'])
        #plot
        if pl==1:
        plt.figure(figsize=(12, 6))
        plt.plot(series.index, series.values, label='Original data', color='indigo')
        plt.plot(forecast_df.index, forecast_df['Forecast'], label='Forecast ARMA + GARCH', color='hotpink')
        plt.fill_between(volatility_df.index, forecast_df['Forecast'] - volatility_df['Volatility'], forecast_df['Forecast'] + volatility_df['Volatility'], color='hotpink', alpha=0.2)
        plt.plot(df[df['anno']==2024]['TMEDIA C']-det,color='violet', label='Observed')
        plt.legend()
        plt.title(f'Forecast for 2024 year - {name}')
        plt.show()
        return forecast_df

def mean_bias_error(y_true, y_pred):
        y_true = np.array(y_true)
        y_pred = np.array(y_pred)
        mbe = np.mean(y_pred - y_true)
        return mbe

def forecast_evaluation(name,df,p,q):
        pepe=forecast(p,q,name,df,1)
        for i in range(20):
        pepo=forecast(p,q,name,df)
        plt.plot(pepo['Forecast']+deterministic[name],color='violet', label='Confidence interval')
        plt.plot(pepe['Forecast']+deterministic[name],color='hotpink', label='Forecast')
        plt.plot(df[df['anno']==2024]['TMEDIA C'],color='indigo',label='
```

```
Original data')
    plt.title(f'Forecast vs observed data - {name} 2024')
    plt.grid(True)
    plt.show()
    #errors
    y_real = np.array(df[(df['anno']==2024)&(df.index!='2024-12-31')]['TMEDIA C'].dropna())
    y_sim = np.array((pepe['Forecast']+deterministic[name]).dropna())

    mae = mean_absolute_error(y_real, y_sim)
    mbe=mean_bias_error(y_real, y_sim)
    corr, _ = pearsonr(y_real, y_sim)
    eror=pd.DataFrame([mae,mbe,corr], index=['MAE','MBD','Pearson'])
    return eror
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker

def cdd_mean_prediction_all(citta_list, nomi_list):
    risultati = []

    fig, axes = plt.subplots(3, 3, figsize=(18, 12))
    axes = axes.flatten()
    fig.suptitle("CDD from May to September   observed vs. historical mean", fontsize=18)


    for i, (citta, nome) in enumerate(zip(citta_list, nomi_list)):
        cdd = []
        for c in range(2005, 2025):
            mask = (citta['anno'] == c) & (citta['mese'].isin([5, 6, 7, 8, 9]))
            temps = citta[mask]['TMEDIA C']
            result = np.maximum(temps - 18, 0)
            cdd.append(result.sum())

    anni = np.arange(2005, 2025)
    valori = np.array(cdd)
    y_train = valori[:19]  # 2005  2023
    valore_2024 = valori[19]
    mean_val = y_train.mean()

    errore_relativo = abs(valore_2024 - mean_val) / abs(mean_val)
    erro_percentuale = errore_relativo * 100
```

```

        ax = axes[i]
        ax.scatter(anni[:19], y_train, color='indigo', label='Observed values (2005 2023)', marker='*', s=80)
        ax.scatter(2024, valore_2024, color='violet', label='Observed value (2024)', marker='*', s=80)
        ax.scatter(2024, mean_val, color='teal', label='Mean value (2005 2023)', marker='*', s=80)

        ax.set_title(f"{nome}")
        ax.set_xlabel("Year")
        ax.set_ylabel("CDD")
        ax.grid(True)
        ax.xaxis.set_major_locator(ticker.MaxNLocator(integer=True))


        risultati.append({
                "City": nome,
                "Observed_2024": valore_2024,
                "Predicted_by_mean": mean_val,
                "Relative_Error": errore_relativo,
                "Percentage_Error": erro_percentuale
        })

        for j in range(i + 1, len(axes)):
                fig.delaxes(axes[j])

        fig.subplots_adjust(bottom=0.80)
        handles, labels = axes[0].get_legend_handles_labels()
        fig.legend(handles, labels, loc='lower center', bbox_to_anchor=(0.5, -0.02),
        ncol=3, fontsize='medium', frameon=False)
        plt.tight_layout()
        plt.show()

        return pd.DataFrame(risultati).set_index("City")

def hdd_mean_prediction_all(citta_list, nomi_list):
        risultati = []

        fig, axes = plt.subplots(3, 3, figsize=(18, 12))
        axes = axes.flatten()  # Rende l'array 2D in 1D per iterare pi facilmente
        fig.suptitle("HDD from November to March  observed vs. historical mean", fontsize=18)


```

```python
        for i, (citta, nome) in enumerate(zip(citta_list, nomi_list)):
                hdd = []
                for c in range(2005, 2025):
                mask = (citta['anno'] == c) & (citta['mese'].isin
                    ([11,12,1, 2, 3]))
                temps = citta[mask]['TMEDIA C']
                result = np.maximum(18 - temps, 0)
                hdd.append(result.sum())

        anni = np.arange(2005, 2025)
        valori = np.array(hdd)
        y_train = valori[:19]  # 2005  2023
        valore_2024 = valori[19]
        mean_val = y_train.mean()

        errore_relativo = abs(valore_2024 - mean_val) / abs(mean_val)
        errore_percentuale = errore_relativo * 100

        ax = axes[i]
        ax.scatter(anni[:19], y_train, color='indigo', label='Observed
            values (2005 2023)', marker='*', s=80)
        ax.scatter(2024, valore_2024, color='violet', label='Observed
            value (2024)', marker='*', s=80)
        ax.scatter(2024, mean_val, color='teal', label='Mean value (2005
             2023)', marker='*', s=80)

        ax.set_title(f"{nome}")
        ax.set_xlabel("Year")
        ax.set_ylabel("HDD")
        ax.grid(True)
        ax.xaxis.set_major_locator(ticker.MaxNLocator(integer=True))

        risultati.append({
                "City": nome,
                "Observed_2024": valore_2024,
                "Predicted_by_mean": mean_val,
                "Relative_Error": errore_relativo,
                "Percentage_Error": errore_percentuale
        })

        for j in range(i + 1, len(axes)):
        fig.delaxes(axes[j])
                fig.subplots_adjust(bottom=0.80)
        handles, labels = axes[0].get_legend_handles_labels()
        fig.legend(handles, labels, loc='lower center', bbox_to_anchor
            =(0.5, -0.02),
        ncol=3, fontsize='medium', frameon=False)
```

```
        plt.tight_layout()
        plt.show()

        return pd.DataFrame(risultati).set_index("City")
```

Listing A.7: Code to compute historical weather indices (HDD or CDD) based on observed temperature data

```
def alaton_sim(nome,citta):
        c=nome
        c2=citta
        sigma_monthly = list(std[c])
        A = params[c]['A']
        B= params[c]['B']
        C = params[c]['C']
        omega = 2 * np.pi / 365
        phi = params[c]['phi']
        x0 = float(c2[c2.index=='2023-10-31']['TMEDIA C'] )

        dati_city = dmedie[c].loc["2005-01-01":"2023-10-31"]
        dati_city = dati_city.to_frame(name='temperature')

        dati_city['t'] = np.arange(1, len(dati_city) + 1)
        dati_city['T_m'] = A + B * dati_city['t'] + C * np.sin(omega *
           dati_city['t'] + phi)
        dati_city['delta_T'] = dati_city['temperature'].diff()
        dati_city['diff_from_Tm'] = -dati_city['temperature'] +
           dati_city['T_m']
        reg_data = dati_city.dropna()
        a_mensile, a_medio = calculate_monthly_a(reg_data)
        som=[]
        for i in range(20000):
                df_temp = simulate_temperature_trajectory(
                start_date="2023-11-01",
                #start_date="2024-01-16",
                end_date="2024-03-31",
                sigma_monthly=sigma_monthly,
                A=A, B=B, C=C, omega=omega, phi=phi, a_mensile=a_mensile
                    , x0=x0
                )
                result = np.maximum(-df_temp +18, 0)
                result_array = result.to_numpy()
                somma= result_array.sum()
                som.append(somma)
        res_somma = np.array(som)
        return res_somma.mean()
```

Listing A.8: Code used to simulate temperature trajectories for Monte Carlo methods based on a modified Ornstein–Uhlenbeck process

```
def simulate_arma_garch_hdd(
        city_name,
        arma_params_df,
        garch_params_df,
        initial_series,
        deterministic_part,
        days_to_simulate=151,
        simulations=20000,
        seed=42
        ):

        #estrazioni parametri ARMA
        arma_row = arma_params_df.loc[city_name].dropna()
        const = arma_row.get('const', 0)
        ar_params = [arma_row.get(f'ar.L{i}', 0) for i in range(1, 4) if f'ar.L{i}' in arma_row]
        ma_params = [arma_row.get(f'ma.L{i}', 0) for i in range(1, 4) if f'ma.L{i}' in arma_row]

        p, q = len(ar_params), len(ma_params)

        #estrazione parametri GARCH
        garch_row = garch_params_df[garch_params_df.index == city_name].iloc[0]
        omega = garch_row['Omega']
        alpha = garch_row['Alpha1']
        beta = garch_row['Beta1']
        nu = garch_row['Nu']

        #serie iniziale
        series = initial_series.dropna().values
        last_vals = list(series[-p:])  # ultimi valori AR
        last_errs = list(np.zeros(q))  # inizializziamo gli errori
        sigma2 = np.var(series[-30:])  # var iniziale

        #parte deterministica
        det = deterministic_part[city_name][-days_to_simulate:].values

        all_sims = []

        for sim in range(simulations):
        sim_vals = []
        sigma2_t = sigma2
```

```
        errs = list(last_errs)
        x_vals = list(last_vals)

        for t in range(days_to_simulate):
        #GARCH volatility
        if len(errs) > 0:
        sigma2_t = omega + alpha * errs[-1]**2 + beta * sigma2_t
        else:
        sigma2_t = omega / (1 - alpha - beta)

        std_t = np.sqrt(sigma2_t)
        eps_t = np.random.standard_t(df=nu) * std_t

        #ARMA forecast
        ar_term = sum([ar_params[i] * x_vals[-(i+1)] for i in range(p)]) if p > 0 else 0
        ma_term = sum([ma_params[j] * errs[-(j+1)] for j in range(q)]) if q > 0 else 0

        x_t = const + ar_term + ma_term + eps_t
        x_vals.append(x_t)
        errs.append(eps_t)
        y_t = x_t + det[t]
        hdd_t = max(18 - y_t, 0)
        sim_vals.append(hdd_t)

        all_sims.append(sum(sim_vals))

        return all_sims
```

Listing A.9: Code used to simulate temperature trajectories for Monte Carlo methods based on ARMA+GARCH process

Chapter B

Additional plots

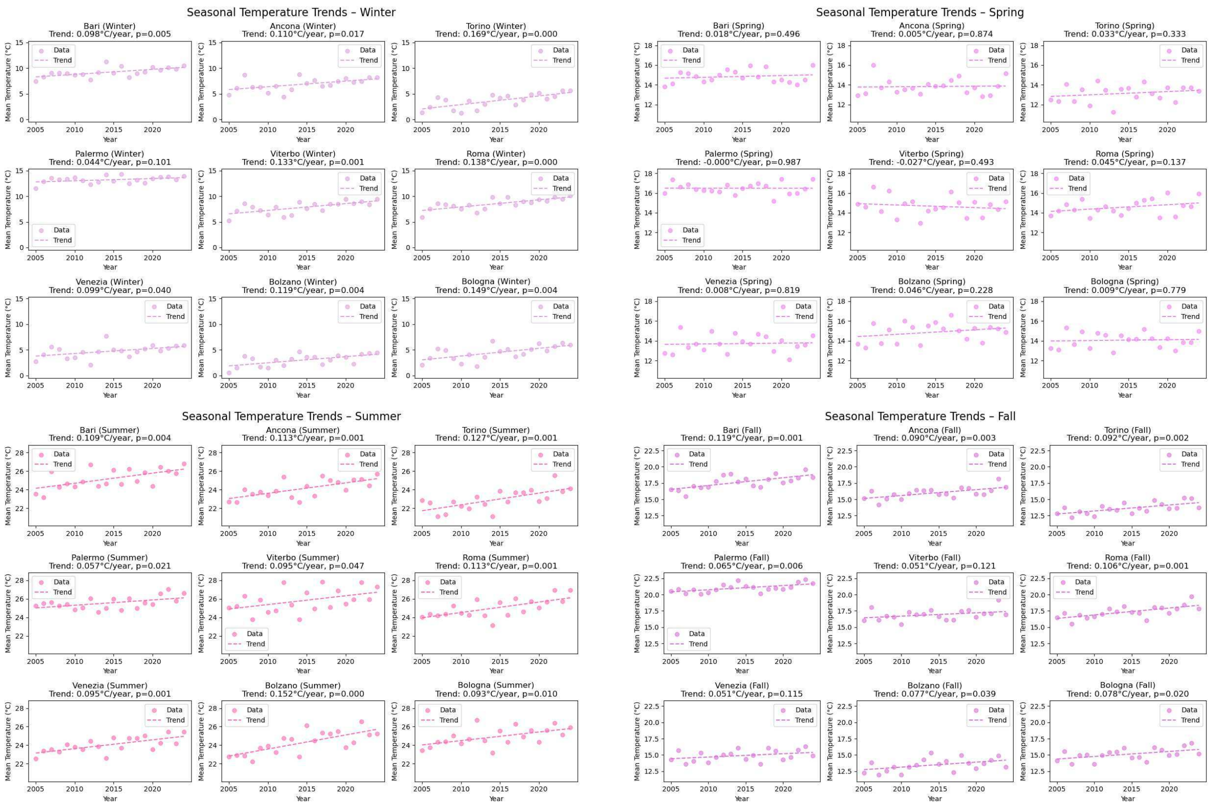


Figure B.1: Seasonal linear trends.

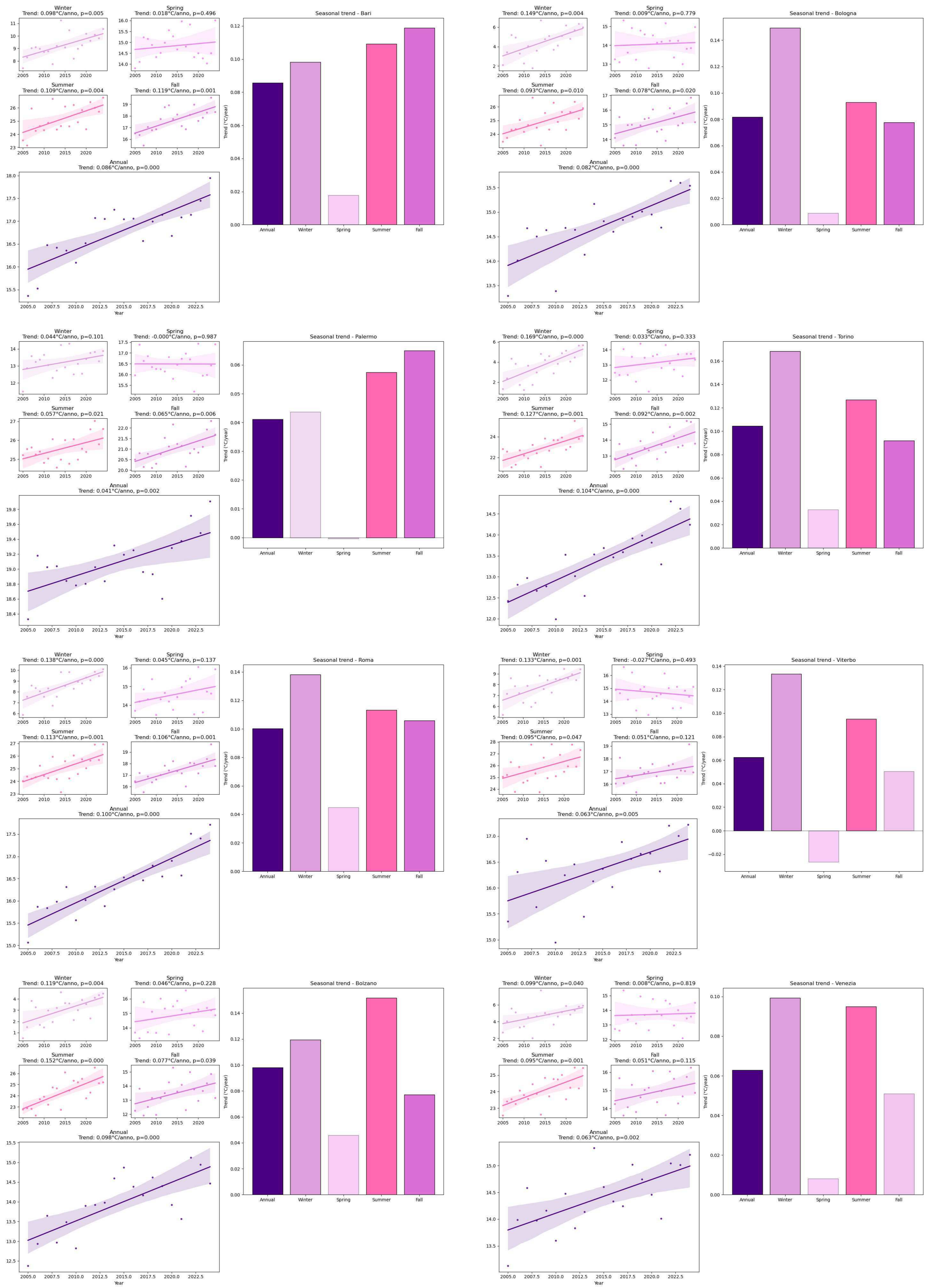


Figure B.2: Linear trend for each city.

Bibliography

[1] Peter Alaton, Boualem Djehiche, David Stillberger, *On modelling and pricing weather derivatives*, Applied Mathematical Finance, 2002, Volume 9, Issue 1, Pages 1–20, Routledge.

[2] Fred Espen Benth and Jurate Saltyte Benth, *Modeling and Pricing in Financial Markets for Weather Derivatives*, World Scientific Publishing Company, 2012.

[3] Stephen Jewson and Anders Brix, *Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations*, Cambridge University Press, 2005.

[4] Rodrigo Caballero, Stephen Jewson and Anders Brix, *Long memory in surface air temperature: detection, modeling, and application to weather derivative valuation*, Clim Res 21:127-140, 2002.

[5] Mengshan Cao and Jiang Wei, *Weather Derivatives Valuation and Market Price of Weather Risk*, Journal of Futures Markets, John Wiley & Sons, Ltd, vol. 24, no. 11, pp. 1065–1089, 2004.

[6] CME Group, *Weather Futures and Options Codes*, `https://www.cmegroup.com/trading/weather/files/updated-weather-product-slate.pdf`, 2024, Accessed: 2025-04-02.

[7] CME Group, *Weather Derivatives Brochure*, `https://www.cmegroup.com/trading/weather/files/WT-124_WeatherBrochure_r11.pdf`, N.D., Accessed: 2025-04-02.

[8] Gregor Dorfleitner, Matthias Wimmer, Sebastian Wimmer, *The Pricing of Temperature Futures at the CME*, The Journal of Risk Finance, Emerald Group Publishing Limited, 2011, Volume 12, Issue 1, Pages 59–74.

[9] Mauro Del Corno, *Derivati sul clima più precisi del meteo e le aziende si assicurano contro il maltempo*, `https://www.ilfattoquotidiano.it/2014/08/26/derivati-sul-clima-piu-precisi-del-meteo-e-le-aziende-si-assicurano-contro-il-maltempo/1099205/`, 2014, Accessed: 2025-04-01.

[10] Paul Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer, 2004.

[11] Stephen Jewson and Anders Brix, *Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations*, Cambridge University Press, 2005.

[12] Peter E. Kloeden and Eckhard Platen, *Numerical Solution of Stochastic Differential Equations*, 3rd edition, Springer, 2011, ISBN: 978-3-642-14452-2.

[13] Tom L. McKnight and Darrel Hess, *Physical Geography: A Landscape Appreciation*, 12th edition, Prentice Hall, 2019.

[14] Yuh-Dauh Lyuu, *Financial Engineering and Computation: Principles, Mathematics, Algorithms*, Cambridge University Press, 2002.

[15] Joseph Lee Rodgers and W. Alan Nicewander, *Thirteen Ways to Look at the Correlation Coefficient*, The American Statistician, 1988, Volume 42, Number 1, Pages 59–66, [American Statistical Association, Taylor & Francis, Ltd.], `http://www.jstor.org/stable/2685263`, Accessed: 2025-05-07.

[16] Mario Pinna, *Climatologia*, UTET, 1977.

[17] Mario Pinna, *Contributo alla classificazione del clima d'Italia*, Rivista Geografica Italiana, 1970, Volume 2, Pages 129–152, Month: Giugno.

[18] I. Colin Prentice, Wolfgang Cramer, Sandy Harrison, Rik Leemans, Robert Monserud, et al., *A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate*, Journal of Biogeography, 1992, Volume 19, Issue 2, Pages 117–134, `https://doi.org/10.2307/2845499`.

[19] Chien-Hsiu Huang and Anna Petukhina, *Applied Time Series Analysis and Forecasting with Python*, Springer International Publishing, 2022.

[20] World Meteorological Organization, *State of the Global Climate 2024*, WMO-No. 1368, World Meteorological Organization, 2025, `https://wmo.int/sites/default/files/2025-03/WMO-1368-2024_en.pdf`, Accessed: 2025-04-22.