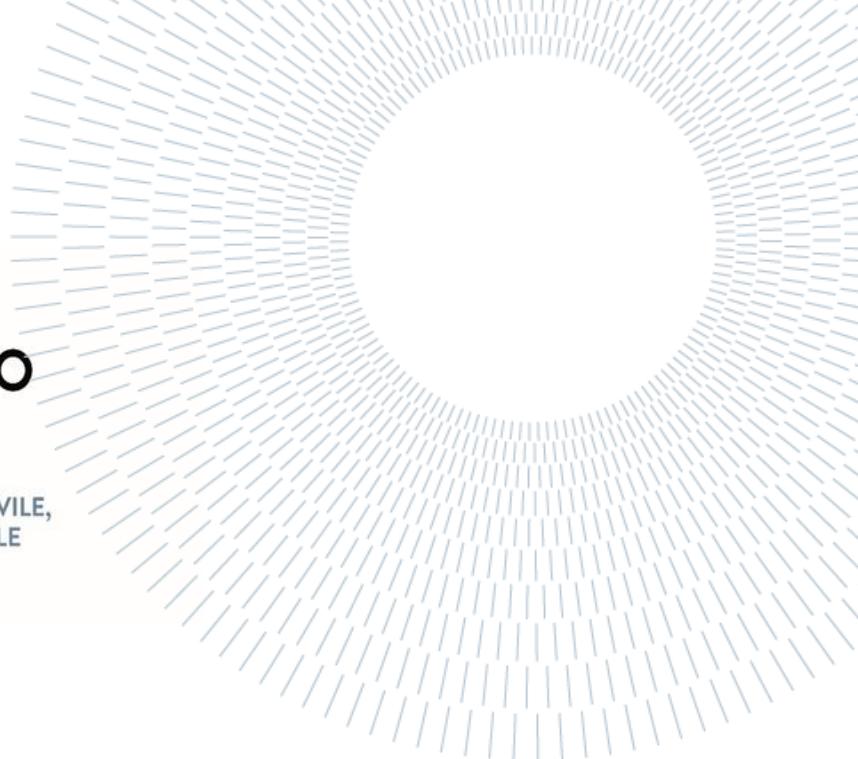




**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA CIVILE,  
AMBIENTALE E TERRITORIALE



# Using Machine Learning to reconstruct temperature and precipitation climatologies in an Italian citizen science weather station network

TESI DI LAUREA MAGISTRALE IN ENVIRONMENTAL AND  
LAND PLANNING ENGINEERING - INGEGNERIA PER  
L'AMBIENTE E IL TERRITORIO

Author: **Francesca Rampinelli**

Student ID: 10503829

Advisor: Prof. Alessandro Ceppi

Co-advisor: Guido Cioni

Academic Year: 2023-24



# Abstract

Awareness of the ongoing changes enables more effective preparation for the challenges of climate change, facilitating the adoption of appropriate mitigation and adaptation strategies. For this reason, climate datasets of meteorological variables have become increasingly important in recent years and are likely to become even more relevant in the future. The objective of this thesis is to develop a methodology for the reconstruction of climatologies for the period 1991-2020 of the Citizen Science Meteonetwork stations, utilising daily data of maximum, minimum, mean temperature and cumulative precipitation. The datasets of NNW stations are incomplete, or lack sufficient observation periods, to allow for the calculation of climate trends over a minimum period of 30 years. The proposed methodology employs machine learning (ML) techniques to establish relationships between the observed data and the reanalysis datasets, thereby enabling the generation of synthetic time series that are conducive to the calculation of climatologies. The process commenced with an initial comparison of various reanalysis datasets to identify the most suitable one for use as input. This was followed by a comparison of two algorithms, Support Vector Regression (SVR) and Random Forest (RF), to determine the optimal algorithm for each variable. Subsequently, the models were optimised for each station and variable, and the series were reconstructed for the period 1991-2020. Ultimately, monthly climatologies were calculated for each variable and station. The results demonstrate that the CERRA dataset and the SVR model with a linear kernel yielded the most optimal results for temperature, whereas MSWEP and SVR with a radial kernel exhibited the highest degree of accuracy for precipitation. The temperature models demonstrated excellent performance, with coefficients of determination exceeding 0.9. In contrast, the precipitation models exhibited considerable variability, with some models failing to achieve the desired level of accuracy. The comparison between the reconstructed and observed climatologies corroborates the efficacy of the methodology for temperature, whereas the results for precipitation are less precise. The methodology shows promise in addressing the scarcity of observational data, but requires refinement, particularly in regard to precipitation, with an enhancement in the quality of data employed for model training.

**Key-words:** climatology reconstruction; support vector regression; random forest; reanalysis datasets; meteorological variables.



## Abstract in italiano

La consapevolezza dei cambiamenti in atto permette di prepararsi in modo più efficace ad affrontare le sfide del cambiamento climatico, favorendo l'adozione di strategie adeguate di mitigazione e adattamento. Per questo motivo, i dataset di climatologie di variabili meteorologiche hanno acquisito crescente importanza negli ultimi anni e diventeranno probabilmente ancora più rilevanti in futuro. L'obiettivo di questa tesi è sviluppare una metodologia per ricostruire le climatologie del periodo 1991-2020 delle stazioni della rete Citizen Science Meteonetwork, utilizzando dati giornalieri di temperatura massima, minima, media e precipitazioni cumulate. Le stazioni presentano lacune nei dati o periodi di osservazione insufficienti per calcolare trend climatici su almeno 30 anni. La metodologia proposta si avvale di tecniche di machine learning (ML) per stabilire relazioni tra i dati osservati e i dataset di rianalisi, consentendo la creazione di serie storiche sintetiche utili al calcolo delle climatologie. Il processo ha inizialmente confrontato diversi dataset di rianalisi per identificare quello più adatto da usare come input per i modelli, seguito da un confronto tra due algoritmi, Support Vector Regression (SVR) e Random Forest (RF), per determinarne il migliore per ciascuna variabile. Successivamente, i modelli sono stati ottimizzati per ogni stazione e variabile e sono state ricostruite le serie per il periodo 1991-2020. Infine, sono state calcolate le climatologie mensili per ogni variabile e stazione. I risultati mostrano che il dataset CERRA e il modello SVR con kernel lineare hanno fornito le migliori prestazioni per la temperatura, mentre MSWEP e SVR con kernel radiale si sono rivelati più accurati per le precipitazioni. I modelli di temperatura hanno dato prestazioni eccellenti con coefficienti di determinazione superiori a 0.9, mentre quelli per le precipitazioni mostrano variazioni significative, con alcuni che non raggiungono una precisione sufficiente. Il confronto tra le climatologie ricostruite e osservate conferma l'efficacia della metodologia per la temperatura, mentre i risultati per le precipitazioni sono meno accurati. La metodologia è promettente per affrontare la mancanza di dati osservativi, ma richiede miglioramenti, soprattutto per la precipitazione, con un aumento nella qualità dei dati usati per calibrare i modelli.

**Parole chiave:** ricostruzione di climatologie; support vector regression; random forest; dataset di rianalisi; variabili meteorologiche.



# Contents

<b>Abstract.....</b>	<b>i</b>
<b>Abstract in italiano .....</b>	<b>iii</b>
<b>Contents.....</b>	<b>v</b>
<b>1    Introduction.....</b>	<b>1</b>
1.1.    State of the Art .....	1
1.2.    Aim of the Study .....	5
1.3.    Reanalysis Models.....	5
1.4.    Machine Learning .....	6
<b>2    Materials and Methods .....</b>	<b>9</b>
2.1.    Study Area.....	9
2.2.    Data .....	10
2.2.1.    Meteonetwork Observation Series .....	10
2.2.2.    Reanalysis Datasets .....	24
2.2.3.    Extraction of Reanalysis Data .....	31
2.3.    Machine Learning Models .....	32
2.3.1.    Support Vector Regression.....	32
2.3.2.    Random Forest.....	34
2.4.    Performance Metrics.....	36
2.4.1.    Taylor Diagram.....	36
2.4.2.    Bias.....	39
2.4.3.    Coefficient of Determination R <sup>2</sup> .....	39
2.5.    Methods .....	40
2.5.1.    Comparison between Reanalysis Datasets and Observation Series .	40
2.5.2.    Machine Learning Model Optimization.....	44
2.5.3.    Model Application and Reconstruction of Climatologies .....	50
<b>3    Results and Discussion .....</b>	<b>53</b>
3.1.    Comparison of Reanalysis Datasets .....	53
3.1.1.    Temperature Bias.....	53
3.1.2.    Taylor Diagrams .....	56
3.2.    Machine Learning Model Performance .....	63

3.3.    Climatologies .....	79
<b>4    Conclusion .....</b>	<b>89</b>
<b>Bibliography .....</b>	<b>97</b>
<b>Appendix .....</b>	<b>101</b>
<b>List of Figures .....</b>	<b>123</b>
<b>List of Tables .....</b>	<b>125</b>
<b>Acknowledgments .....</b>	<b>127</b>

# 1 Introduction

## 1.1. State of the Art

It is widely acknowledged that the climate is undergoing a rapid and alarming transformation. Anthropogenic global warming has already warmed the planet: from the preindustrial period (1850-1900) to the present (2011-2020) the global average temperature over land has increased by 1.1°C [1]. If global warming continues at the current rate, it is projected that the global average temperature will reach a 1.5°C increase between 2030 and 2052 [1].

As climate change has accelerated, the extent of damage due to abnormal weather phenomena is increasing and becoming more difficult to predict [2]. Consequently, there is a strong need for accurate and quantitative climate data based on meteorological observations [2]. The knowledge of parameters such as average temperature, rainfall and the distribution of extreme events both in the present and in the past represents a significant advance in gaining insight into events which may be related to global warming.

The availability of meteorological data over extended periods enables the calculation of climatologies, defined as the long-term average of a given variable over a period of at least 30 years. Climatologies or climatological normals are fundamental to compare currently observed meteorological variables, such as temperatures and precipitation, with historical averages, thus enabling the detection of any anomalies.

Awareness of the ongoing changes allows for more effective preparation to address the challenges posed by climate change and to implement appropriate strategies for

mitigation and adaptation. For these reasons, datasets of monthly climatological normals of meteorological variables (or climatologies) at a high spatial resolution have proved to be of increasing importance in the recent past, and they are likely to become even more important in the near future [3]. Indeed, they are crucial in a variety of models and decision-supporting tools in a wide spectrum of fields such as agriculture, engineering, hydrology, ecology and natural resource conservation [3].

The Mediterranean region has warmed and will continue to warm more than the global average, particularly in summer [1]. This region will become drier due to the combined effect of decreased precipitation and increased evapotranspiration. At the same time, extreme precipitation will increase in some areas [1]. Italy, situated at the heart of the Mediterranean, is particularly susceptible to the effects of these changes.

In recent years, Italy, like other countries in the Mediterranean area, has been observing an increase in extreme weather events (i.e., heavy rainfall and increase in temperature), causing huge impacts (i.e., floods, droughts, heat waves) with consequences to assets and people [4]. These occurrences are specifically related to peculiar characteristics of the region, such as demographic vulnerability (i.e., high-density population), orographic features, and the complex terrain close to the coast [5]. Moreover, their occurrence in multiple climatic impact-drivers perspectives is expected to increase (with high confidence) since global warming is expected to increase faster over this area than the global mean temperature change [1]. Therefore, there is an increasing need for more reliable and detailed climate information on the Italian peninsula to improve the assessment of climate hazards and risk management [6].

Two recent studies have reconstructed high-resolution climatologies datasets for the entire Italian territory: Brunetti et al. (2014) [7] and Crespi et al. (2018) [8] for temperature and precipitation, respectively. These studies have provided monthly climatologies at a 30-arc-second resolution, based on a high-quality observational

dataset, for the period 1961-1990, using interpolation methods. High-density observational datasets must be integrated with interpolation methods to account for all major factors affecting spatial patterns and to produce reliable estimates even in areas with complex topography or in remote regions such as mountain areas with sparse station coverage [9].

The aforementioned studies represent some of the most comprehensive climatological datasets currently available for Italy. However, they highlight that the greatest challenge lies in obtaining data from the various observational sources and integrating them in a way that ensures consistency across the entire area of interest. One of the main challenges associated with the utilisation of meteorological stations pertains to the presence of incomplete datasets and discrepancies in station characteristics. In fact, in order to construct reliable climatologies, it is essential to have complete and homogeneous time series. However, not all stations inside a network have complete data for the requested period, necessitating the utilisation of interpolations derived from the data of neighbouring stations to address the gaps. The absence of data in certain stations may be attributed to a number of factors, including changes in location or observation methodology. For instance, datasets may originate from meteorological stations that have been relocated over time or that have employed various observation standards. Such discrepancies may result in inhomogeneities within the data, necessitating correction to prevent erroneous reconstructions of the climatology. Moreover, the combination of data from neighbouring stations or from the same station but with disparate characteristics represents an additional challenge. Even minor shifts in the location of a station, particularly in regions characterised by complex orography, can have a significant impact on the measurements, further complicating the process of data homogenisation and interpolation [7].

One potential avenue for enhancing the quantity of accessible data is to incorporate data from citizen science networks, in addition to official national and regional observation networks, in order to increase the geographical coverage of the territory.

In recent decades, thanks to the constant development of information and communication technology (ICT) and the reduction in device costs, interest in meteorology among the general public has grown significantly, leading to the emergence of citizen scientists capable of creating large amateur networks with increasingly reliable automatic measuring stations [10]. Citizen science is now recognized as a significant activity that supports, and often complements, research and institutional monitoring conducted by universities, research centres, and environmental control agencies [11].

Given that a decline in operating ground-based weather stations has been observed in recent years in most countries around the world [12], and that the demand for real-time, high space-time resolution data is increasing, the importance of crowdsourcing weather data becomes clear. As computing power increases, the ability to process and use these kinds of data will also rise; therefore, it is necessary to explore their potential [10].

The Meteonetwork (MNW) system is a prime example of a citizen weather station network (CWS), covering a wide territory with high spatial density, which allows for a high redundancy of measurements. Composed mainly of atmospheric science enthusiasts, MNW has built a vast meteorological network in Italy and Europe, displaying real-time observations with more than 6,500 stations, of which around 3,400 are constantly online during the day [10].

From the MNW experience, it is clear that crowdsourcing has become a valuable tool for both public engagement and scientific research. If appropriate validation and quality control procedures are adopted, citizen science has enormous potential for

providing an additional valuable source of high temporal and spatial resolution real-time data, especially in regions where few observations currently exist [10].

## 1.2. Aim of the Study

This study aims to develop a methodology for reconstructing climatologies for the period 1991-2020 using daily data of maximum, minimum, and average temperatures, as well as cumulative precipitation, from the meteorological stations of the Citizen Science (CS) network Meteonetwork (MNW).

Given that the MNW association was established in 2002, there is a deficit of data from its station network prior to that year. In order to reconstruct climatologies, which require a minimum of 30 years of data, it is necessary to identify a method for obtaining the missing data.

The idea is to utilise the nearest grid points obtained from a gridded reanalysis model to establish a relationship between reanalysis and MNW station data over a specified period (reference), through machine learning (ML) techniques. This represents the first instance of ML models being applied to the MNW network with the specific objective of reconstructing climatologies. Once the relationship has been identified, it can be employed to reconstruct an equivalent synthetic time series for each station over the climatological period 1991-2020, which is the 30-year reference period established by the World Meteorological Organization (WMO).

## 1.3. Reanalysis Models

Reanalysis models currently represent the most comprehensive tool for reconstructing the history of the atmospheric state globally. Such models provide an accurate description of the past weather by assimilating observations from several sources into the most advanced numerical weather prediction models that obey physical laws, in order to create a complete and consistent climate dataset that covers several decades back in time and that is coherent both from physical and dynamic perspectives [13].

Reanalysis does not have the constraint of issuing timely forecasts, so there is more time to collect observations, and when going further back in time, to allow for the ingestion of improved, reprocessed versions of the original observations. By doing so, all available observations will be used and, if possible, with better quality than the original ones.[15]. The model employed for reanalysis is identical to that utilised for forecasting. However, it has been fixed at a past version, thus enabling a more consistent comparison over time.

The use of reanalysis has brought advantages in various sectors, primarily in meteorology and forecasting, where it has been crucial for understanding global climate trends, particularly in recent years. Two aspects have been favouring the application of reanalysis data: (i) their consistency in space and time, covering several decades at global or regional scales; (ii) their free public availability by means of dedicated web platforms, which make these data ready to be used in standard formats, thus avoiding all those time-consuming procedures required for collecting and homogenizing weather station data from different service providers [16].

At present, a number of distinct reanalysis datasets are accessible, exhibiting variability in terms of their spatial domain (global or regional), design objective, data sources (gauge, ground radar, satellite, etc.), bias correction, downscaling approach, variables included, spatiotemporal resolution, temporal coverage, and near-real-time availability. In the Chapters 2 and 3 will be presented and analysed a number of reanalysis datasets in order to identify the most appropriate one for the purposes of this study.

## 1.4. Machine Learning

The application of Machine Learning (ML) techniques represents a promising approach for the effective modelling and reconstruction of climate data. ML is a field of artificial intelligence that employs algorithms to discern and comprehend the concealed relationships between data and information. The popularity of machine

learning is growing rapidly in many areas of human life due to its capacity to process vast quantities of data, solve intricate problems and identify patterns in behaviour.

Machine Learning functions by generalising the training experience and producing a hypothesis that estimates the target function. This capacity for generalization enables the system to make accurate predictions even when presented with data that has not been previously encountered. In contrast to other optimisation problems, machine learning does not have a precise function to optimise. Instead, training errors are employed to facilitate continuous improvement of the learning process.

In the field of hydrology, ML has gained increasing attention because of its learning capability to deal with the nonlinearity and complex problems [17].

A considerable number of studies have demonstrated the effectiveness of ML in improving the accuracy of precipitation and temperature estimation through data integration. Baez-Villanueva et al. (2020) [18] employed a Random Forest (RF) algorithm, to integrate various precipitation data sources with observational information. Kumar et al. (2019) [19] evaluated the skill of Support Vector Regression (SVR), another ML algorithm, by considering the effect of soil moisture on improving the performance of near real-time precipitation. Also in a recent study, Nguyen et al. (2023) [20] investigated the potential of six machine learning algorithms for merging multi-source precipitation products with ground-based data with the aim of enhancing the accuracy of precipitation estimates. The findings of this study indicate that ML approaches can be effectively employed to merge satellite and model-based precipitation products with observed data to improve rainfall estimation accuracy, especially in areas with sparse rain gauge coverage. Two models in particular demonstrated optimal performance: Random Forest (RF) algorithm exhibited one of the most accurate performances, while the Support Vector Regression (SVR) algorithm demonstrated a superior capacity for generalization, particularly in scenarios with low rain gauge station density.

Lee et al. (2018) [2] applied SVR to predict values at weather stations using data from surrounding stations, thereby enabling the identification of anomalies through a comparison of predicted and observed values. This approach was utilised for a range of meteorological variables, including temperature and precipitation. The SVR model was able to capture the geographic and climatic characteristics of observation stations by learning from past data, enabling it to create customized models for each station and meteorological element, giving it an advantage over non-machine learning approaches.

In light of the favourable outcomes observed in the aforementioned studies, this analysis will assess the efficacy of two distinct modelling approaches: Support Vector Regression (SVR) and Random Forest (RF). The Chapters 2 and 3 will provide a detailed account of these models and an evaluation of their performance.

In conclusion, the objective of this study is to contribute to the understanding of climate change in Italy, through the reconstruction of climatologies of temperature and precipitation data from Citizen Science networks, through reanalysis models and ML models. The following chapters will provide a comprehensive account of the methodology employed for the selection of the input reanalysis, the application of Machine Learning techniques, and the validation of the resulting data. These steps are essential to guarantee the accuracy and reliability of the reconstructed climatologies, thereby contributing to an increase in the quantity and the availability of climatological data in Italy and, consequently, to the monitoring of climate change in Italy.

## 2 Materials and Methods

The following chapter describes the data and the methods employed in the analysis, that was structured according to the following workflow:

1. The first step was to compare different reanalysis datasets in order to identify the optimal one to use as input in the Machine Learning procedure;
2. The second step was to compare the two different ML algorithms, SVR and RF, to determine the most suitable model for each one of the following variables: maximum temperature, average temperature, minimum temperature and cumulative precipitation;
3. The next step was to determine the best achievable performance of the model by optimising the model parameters for each station and variable;
4. Subsequently, a synthetic observation series for each station was constructed in the period 1991-2020 using the optimal model;
5. Lastly, the monthly climatologies of each variable of the observation stations was calculated.

All data processing was conducted using the *Python* programming language, with the utilisation of a number of libraries, including *Pandas*, *Sklearn*, *Skillmetrics* and *Matplotlib*.

### 2.1. Study Area

The geographical area of study is Italy, situated in the centre of the Mediterranean region, with an area of approximately 302,000 km<sup>2</sup>. Italy is characterised by a considerable climatic and morphological variety. The country's geographical position,

coupled with the presence of mountain ranges such as the Alps and the Apennines, gives rise to considerable climatic differences between the various regions. The climate varies considerably across Italy, from the temperate continental climate of the northern areas, with cold winters and hot summers, to the Mediterranean climate of the southern coastal areas and island, characterised by hot, dry summers and mild, rainy winters. In contrast, mountainous regions experience lower temperatures and precipitation during the winter snowfall season. The diverse topography of Italy, comprising plains, hills, and mountain ranges, contributes to the country's overall climatic diversity, as evidenced by the weather data collected by the numerous MNW stations distributed across the country.

## 2.2. Data

### 2.2.1. Meteonetwork Observation Series

Meteonetwork (MNW) is an association, born in 2002, and mainly composed by atmospheric science enthusiasts, who built up in time the largest weather station network in Italy. The MNW association has developed an integrated and coordinated network of Automatic Weather Stations (AWS) composed of various amateur meteorological sites. The MNW operational area covers the entire European region, with a particular focus on Italy. The network comprised, in 2022, a total of 6,506 weather stations from 42 countries worldwide. A detailed overview of the MNW network is available on the website <https://meteonetwork.eu/en> (accessed on 1st September 2024).

The MNW network includes both Personal Weather Stations (PWS) and official meteorological networks. Over the past decade, two major mutual data exchange agreements have been established with European amateur associations: Infoclimat in France and MeteoGR in Greece. Moreover, since 2020, additional data have been imported from the Frost network, managed by the Norwegian Meteorological

Institute, which includes quality-controlled daily, monthly, and yearly measurements of temperature, precipitation, and wind data [10].

Locally, mutual data exchange agreements have been signed to automatically integrate data from the Regional Environmental Protection Agencies (ARPA) of some Italian regions, such as Emilia-Romagna, Veneto, and Calabria, and from the meteorological agencies of the autonomous provinces of Trento and Bolzano. In 2021, MNW stations were included in the MISTRAL portal (Meteo Italian SupercompuTing poRtAL, <https://www.mistralportal.it/> accessed on 1st September 2024); this led to a two-way relationship and mutual exchange of weather station data within both networks [10].

Among the 6,506 weather stations currently available, 4,411 are located in Italy, distributed to cover the whole Italian territory. The stations are generally instruments with a good quality/price ratio, such as Davis (Vantage Pro 2 and Vue versions), Ecowitt, Froggit, Sainlogic, Oregon Scientific, Bresser, PCE, Irox, and Lacrosse. These stations mainly measure air temperature, relative humidity, precipitation, wind speed and direction, and atmospheric pressure, although not all of them are equipped with instruments to monitor atmospheric variables such as incoming solar radiation or UV radiation [10].

Of the 6,506 stations, 4,780 upload their data to the MNW database at least once every 24 hours, and about 3,400 are constantly online throughout the day. The sampling frequency for each instantaneous datum ranges from a minimum of 1 minute up to 60 minutes, with most stations having a sampling frequency between 3 and 5 minutes.

Many associations and research organizations adopt installation rules to ensure a standard methodology in data collection. MNW has established its own guidelines based on World Meteorological Organization (WMO) rules, but adapted to the context of Personal Weather Stations (PWS), often located in various types of sites, such as roofs, gardens, or other surfaces. Since applying appropriate quality control methods

is an essential component when using these observations, specific guidelines, standards, and protocols have been established to ensure a high level of trustworthiness, robustness, and to quantify the reliability of crowdsourced data [10].

The admission procedure for MNW stations ensures a certain level of data quality for each sensor, as erroneous data readings caused by sensor malfunctions, hardware or power supply errors, or changed environmental conditions can still occur. MNW data are subjected to automatic validation and quality control procedures to reduce the possibility of measurement errors [10].

The quality control process is divided into two main phases:

1. **Range Test:** Eliminate clearly incorrect data (e.g., minimum temperatures below -40 °C, maximum temperatures above +50 °C, and rainfall below 0 mm).
2. **Cross-validation through percentiles:** For each atmospheric variable and each station, the distribution of values measured by the 15 nearest stations within a maximum range of 30 km is analysed to obtain the 10th and 90th percentiles. A tolerance is calculated using the standard deviation of the distribution, which is added to the 90th percentile and subtracted from the 10th percentile to obtain the cut-off values of the quality check. The use of standard deviation for tolerance computation takes into account the natural variability of the variable in a specific weather situation. If the variability of the given variable is predominantly accentuated (high standard deviation), the boundaries are more relaxed and less stringent. In situations where the natural variability of the variable field is very reduced, the tolerance is more rigorous. A general check over the whole Italian domain with the same logic is performed only for the interpolated-map making process, to be aware of deleting those outliers that might cause graphical issues. Data failing the quality control are flagged and do

not enter the maps elaboration process; however, they remain in the database [10].

The high quality of MNW data has been highlighted in Giazz et al. (2022) [10], where temperature and precipitation data were compared with those from the official networks of ARPA Emilia-Romagna and Veneto. The results underlined the effectiveness of the MNW network and, above all, how the joint use of the two networks has provided a greater or, at worst, null benefit compared to the use of ARPA networks alone [10].

Of the 4,411 meteorological stations comprising the MNW network in Italy, 2,163 are managed by citizen scientists. A total of 43 stations were selected from the latter group to serve as test cases for the methodology proposed in this study.

The selection was based on two principal criteria:

1. The **length of the historical series** was taken into account when selecting the stations. Those with the longest series were chosen, this guarantees the requisite robustness for the precise calibration of the ML model;
2. In terms of **geographical distribution**, the stations were selected with the objective of covering as much of the Italian territory as possible. To guarantee the comprehensive representation of climatic contexts, the study included stations situated in diverse geographical locations, including mountainous and flat areas, as well as those in proximity to and distant from the sea.

The selected stations and their respective characteristics are presented in Table 1, while their location is illustrated in Figure 1. As can be observed, the majority of the selected stations are Davis Vantage models, only one station in Trentino-Alto Adige is Barani. The insertion dates of the station in the MNW database span the period from 2002 to 2012. Also, the altitudes of the stations are highly variable, with the dataset including four mountain stations exceeding 1,000 meters: *trn017* (2,450 meters), *vda006* (1,500

meters), *lmb255* (1,386 meters), and *ero001* (1,118 meters). In the intermediate hilly zone, four stations are situated at elevations between 750 and 595 meters above sea level. These are *lmb002* (750 m), *pmn036* (686 m), *abr034* (685 m), and *tsc002* (595 m). The remaining stations are all located at an altitude below 500 meters.

Table 1: Selected MNW observation stations and their characteristics.

Code	Elevation [m]	Insertion Date	Strumentation	Location	Region
abr034	685	2011-05-15	Davis Vantage Pro 2	L'Aquila (AQ)	Abruzzo
abr047	85	2012-11-03	Davis Vantage Pro 2 Plus	Pescara (PE)	Abruzzo
bsl008	455	2008-03-29	Davis Vantage Pro 2	Miglionico (MT)	Basilicata
clb002	173	2005-05-17	Davis Vantage Pro 2	Rende (CS)	Calabria
cmp015	69	2009-05-11	Davis Vantage Pro 2	Pignataro Maggiore (CE)	Campania
ero001	1118	2005-05-15	Davis vantage Pro 2	Frassinoro (MO)	Emilia-Romagna
ero002	31	2002-11-01	Davis Vantage Pro	Gambettola (FC)	Emilia-Romagna
ero009	38	2003-08-26	Davis Vantage Pro	Anzola dell'Emilia (BO)	Emilia-Romagna
fvg009	35	2007-11-12	Davis Vantage Pro 2	Fontanafredda (PN)	Friuli-Venezia Giulia
laz013	13	2006-02-23	Davis Vantage Pro 2	Latina (LT)	Lazio
laz026	320	2006-11-13	Davis Vantage Vue	Marta (VT)	Lazio
laz033	362	2007-12-10	Davis Vantage Pro 2	Ferentino (FR)	Lazio
lig009	130	2008-02-28	Davis Vantage Pro 2	Zoagli (GE)	Liguria
lmb002	750	2002-11-01	Davis Vantage Pro 2	Barzio (LC)	Lombardia
lmb015	28	2002-11-01	Davis Vantage Vue	Mantova (MN)	Lombardia
lmb021	224	2002-11-01	Davis Vantage Pro 2	Seregno (MB)	Lombardia
lmb039	68	2002-11-23	Davis Vantage Pro	Codogno (LO)	Lombardia
lmb084	427	2004-12-05	Davis Vantage Pro 2	Capiago Intimiano (CO)	Lombardia

Code	Elevation [m]	Insertion Date	Strumentation	Location	Region
lmb255	1386	2011-07-05	Davis Vantage Pro 2	Valfurva (SO)	Lombardia
mbr001	455	2002-01-01	Davis Vantage Pro 2	Gualdo Tadino (PG)	Umbria
mbr006	300	2005-09-17	Davis Vantage Pro 2	Marsciano (PG)	Umbria
mbr026	350	2012-07-09	Davis Vantage Pro	Orvieto (TR)	Umbria
mrc009	25	2005-10-27	Davis Vantage Pro 2	Grottammare (AP)	Marche
pgl005	195	2005-01-21	Davis Vantage Pro 2	Terlizzi (BA)	Puglia
pgl008	76	2005-05-16	Davis Vantage Pro 2	Foggia (FG)	Puglia
pgl012	106	2006-10-23	Davis Vantage Pro 2	Supersano (LE)	Puglia
pgl040	20	2009-10-16	Davis Vantage Pro 2	Vieste (FG)	Puglia
pmn016	229	2003-12-29	Davis Vantage Pro 2	Nichelino (TO)	Piemonte
pmn033	170	2006-09-23	Davis Vantage Pro 2	Vicolungo (NO)	Piemonte
pmn036	686	2006-11-13	Davis Vantage Pro 2	Pila (VC)	Piemonte
pmn074	200	2010-10-27	Davis Vantage Pro 2	Pietra Marazzi (AL)	Piemonte
scl007	430	2004-09-25	Davis Vantage Pro	Partanna (TP)	Sicilia
scl014	170	2005-05-16	Davis Vantage Pro	Vittoria (RG)	Sicilia
scl073	95	2011-03-28	Davis Vantage Vue	Termini Imerese (PA)	Sicilia
sdr013	339	2010-03-04	Davis Vantage Pro 2	Ossi (SS)	Sardegna
trn009	202	2005-12-20	Barani MeteoHelix IoT Pro	Trento (TN)	Trentino-Alto Adige
trn017	2450	2007-04-03	Davis Vantage Pro	Ortisei (BZ)	Trentino-Alto Adige
tsc001	2	2002-11-01	Davis Vantage Pro 2 Plus	Pisa (PI)	Toscana

Code	Elevation [m]	Insertion Date	Strumentation	Location	Region
tsc002	595	2002-11-01	Davis Vantage Pro 2	Castelnuovo di Val di Cecina (PI)	Toscana
tsc003	60	2002-11-01	Davis Vantage Pro	Prato (PO)	Toscana
vda006	1500	2012-08-22	Davis Vantage Pro 2	Ollomont (AO)	Valle d'Aosta
vnt023	90	2005-06-11	Davis Vantage Pro	Breganze (VI)	Veneto
vnt160	2	2010-11-04	Davis Vantage Pro 2	Cavallino-Treporti (VE)	Veneto

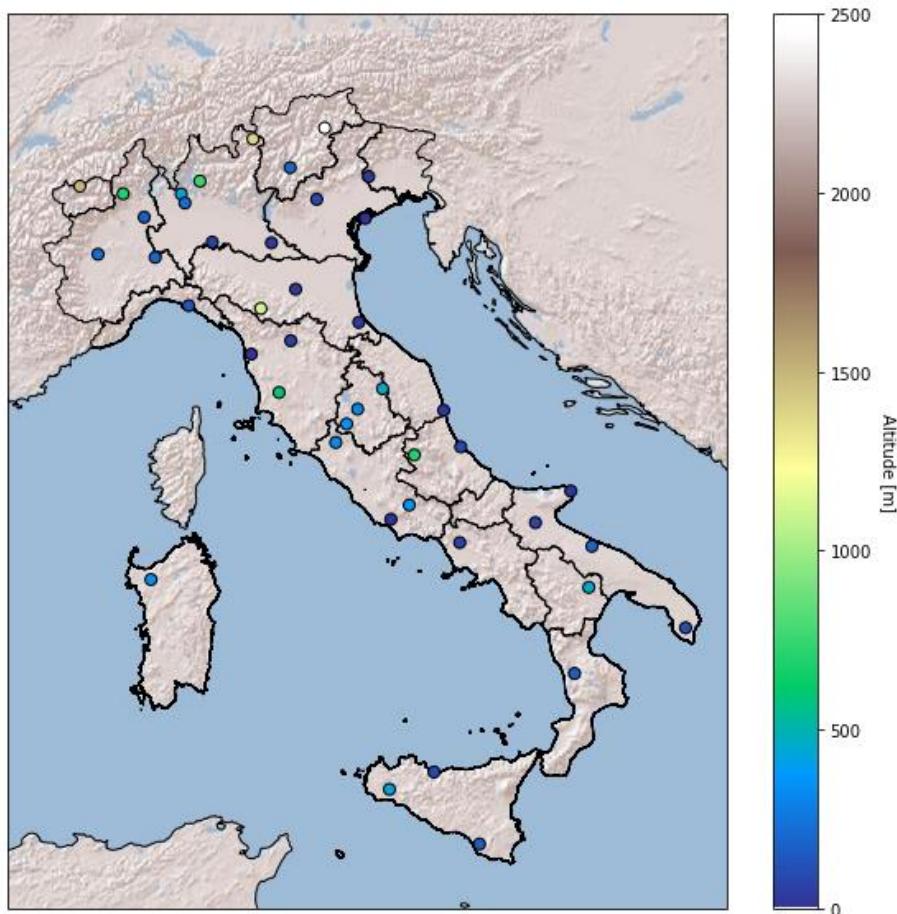


Figure 1: Selected MNW observation station, their location and altitude in meters.

Subsequent to the selection of the stations, the complete historical archive of daily data of each station was downloaded from the MNW database, the last occasion upon

which data was downloaded from the MNW database was in October 2023. The downloaded historical series included the following variables:

- **Daily maximum temperatures (at 2 metres)**, which represent the actual daily maximum extremes recorded by the station, rather than simply the maximum temperature observed over the course of an hour;
- **Daily average temperatures (at 2 metres)**, which are the mean of all daily recorded data, not only hourly data;
- **Daily minimum temperatures (at 2 metres)**, which represent as well the actual daily minimum extremes recorded by the station;
- **Daily cumulative precipitation**, that includes both solid and liquid sources, but as the majority of rain gauges are not heated during the winter months, this may result in some discrepancies. This could result in an underestimation of precipitation in the form of snow.

Table 1 shows that the stations were inserted in the MNW database at disparate times between 2002 and 2012, resulting in the observation time series for each station commencing at varying points in time. Consequently, the length of the observation periods varies by up to ten years between stations.

Even though all the downloaded series were subjected to a validation process to ensure the reliability of the observations, it is important to note that, despite the emphasis on data quality, some series may have incomplete data. The absence of daily observations may be attributed to transient issues with the measuring instrument or temporary deficiencies in the observation recording and storage processes. It was essential therefore, to conduct a comprehensive analysis of the observation periods for each station in order to ascertain the extent of missing data and the effective length of each historical time series.

To this end, as a preliminary step, the presence of unavailable values (NaN) in the dataset was meticulously examined to guarantee that subsequent analyses are conducted on a comprehensive and dependable dataset. Rows containing missing values were removed from the dataset. If a precipitation observation occurs on a given date but no temperature observation was recorded, or vice versa, the corresponding row was retained only for the processing of the present variable. Indeed, for the variables for which no observation is available, the row was excluded.

Subsequently, the exact locations of missing values were identified, namely where there was a gap between dates of observation and the row was actually absent. This facilitated a more comprehensive understanding of the distribution of gaps in the dataset. For each variable and weather station, continuous observation periods were identified, defined as sequences of consecutive days without gaps in the data. The identification was conducted through the utilisation of a dedicated function that enables the aggregation of data by dates, including the calculation of the duration of each continuous period and the delineation of the extent of the gaps.

Consequently, the total number of observation days and the number of days with missing data for each station and each variable were calculated. In order to calculate the number of missing observations, the difference between the total number of days in each year and month and the present days was made, with the impact of leap years also considered. Table 2 and Table 3 present an illustrative example, showcasing the present and absent days of observation for the variables mean temperature and precipitation.

Table 2: Total number of observation days and days with missing data for each station for the variable mean temperature.

Code	Total Observation Days	Total missing Days	Total Observation Years	Total missing Years
abr034	4395	121	12	0.3
abr047	3777	215	10.3	0.6

Code	Total Observation Days	Total missing Days	Total Observation Years	Total missing Years
bsl008	5467	205	15	0.6
clb002	4924	1111	13.5	3
cmp015	4876	388	13.3	1.1
ero001	6150	562	16.8	1.5
ero002	6911	724	18.9	2
ero009	3842	2175	10.5	6
fgv009	5536	273	15.2	0.7
laz013	6187	250	16.9	0.7
laz026	5045	1131	13.8	3.1
laz033	5522	177	15.1	0.5
lig009	4880	125	13.4	0.3
lmb002	7383	254	20.2	0.7
lmb015	6326	812	17.3	2.2
lmb021	8288	40	22.7	0.1
lmb039	6688	937	18.3	2.6
lmb084	6761	132	18.5	0.4
lmb255	4347	144	11.9	0.4
mbr001	4740	2551	13	7
mbr006	5852	145	16	0.4
mbr026	3731	19	10.2	0.1
mrc009	4768	532	13.1	1.5
pgl005	5834	1001	16	2.7
pgl008	5345	1375	14.6	3.8
pgl012	4736	1105	13	3
pgl040	4768	332	13.1	0.9
pmn016	7007	217	19.2	0.6
pmn033	6222	3	17	0
pmn036	6074	15	16.6	0
pmn074	4060	380	11.1	1
scl007	6773	113	18.5	0.3
scl014	6190	530	16.9	1.5
scl073	4086	260	11.2	0.7
sdr013	3882	1085	10.6	3
trn009	6010	492	16.5	1.3
trn017	5001	1032	13.7	2.8
tsc001	7579	58	20.8	0.2

Code	Total Observation Days	Total missing Days	Total Observation Years	Total missing Years
tsc002	7557	75	20.7	0.2
tsc003	7054	407	19.3	1.1
vda006	3404	200	9.3	0.5
vnt023	6236	458	17.1	1.3
vnt160	4309	56	11.8	0.2

Table 3: Total number of observation days and days with missing data for each station for the variable precipitation.

Code	Total Observation Days	Total missing Days	Total Observation Years	Total missing Years
abr034	4393	123	12	0.3
abr047	3777	215	10.3	0.6
bsl008	5466	206	15	0.6
clb002	4924	1111	13.5	3
cmp015	4876	388	13.3	1.1
ero001	6150	562	16.8	1.5
ero002	6911	724	18.9	2
ero009	3842	2175	10.5	6
fgv009	5536	273	15.2	0.7
laz013	6187	250	16.9	0.7
laz026	5045	1131	13.8	3.1
laz033	5522	177	15.1	0.5
lig009	4880	125	13.4	0.3
lmb002	7383	254	20.2	0.7
lmb015	6326	812	17.3	2.2
lmb021	8316	12	22.8	0
lmb039	6688	937	18.3	2.6
lmb084	6761	132	18.5	0.4
lmb255	4347	144	11.9	0.4
mbr001	4740	2551	13	7
mbr006	6574	24	18	0.1
mbr026	3731	19	10.2	0.1
mrc009	4768	532	13.1	1.5
pgl005	5834	1001	16	2.7

Code	Total Observation Days	Total missing Days	Total Observation Years	Total missing Years
pgl008	5345	1375	14.6	3.8
pgl012	4736	1105	13	3
pgl040	4768	332	13.1	0.9
pmn016	7007	217	19.2	0.6
pmn033	6222	3	17	0
pmn036	6074	15	16.6	0
pmn074	4060	380	11.1	1
scl007	6773	113	18.5	0.3
scl014	6190	530	16.9	1.5
scl073	4086	260	11.2	0.7
sdr013	3882	1085	10.6	3
trn009	6010	492	16.5	1.3
trn017	5000	1033	13.7	2.8
tsc001	7579	58	20.8	0.2
tsc002	7557	75	20.7	0.2
tsc003	7054	407	19.3	1.1
vda006	3408	196	9.3	0.5
vnt023	6245	449	17.1	1.2
vnt160	4309	56	11.8	0.2

Lastly, a bar graph was constructed for each variable, illustrating the observation periods for each station and highlighting the temporal position of missing data. Furthermore, in the graph was added the effective duration of observations series in years, with only observed days included and missing days excluded.

As illustrated in Figures 2-5, the observation periods of several stations exhibit a relatively continuous temporal coverage, demonstrating a high level of consistence in the temporal resolution of the data. In contrast, some stations, such as *mbr001* or *pgl008*, exhibit a significant number of gaps, which reflect substantial interruptions in the observational process. The graphs indicate that numerous stations exhibit common gaps for all the four variables, which may be attributed to external factors such as instrumental malfunctions or data collection issues. In accordance the observation

periods of stations with continuous data for one variable also tend to exhibit similar trends for the others. This could indicate effective overall station management or favourable local conditions for data collection. Indeed, it is typical for instances of missing temperature data to be accompanied by instances of missing precipitation data, and vice versa. This explains why the observation periods for each variable are similar for the majority of stations, differing by only a few days.

Furthermore, it is notable that all but one of the selected stations have at least a decade's worth of observations for each variable. *Vda006* is the single station with a time series of just over nine years for all four variables. Four stations have observation series of more than 20 years (*lmb002*, *lmb021*, *tsc001*, *tsc002*). The station with the longest time series is *lmb021*, with a period of approximately 23 years for all the variables.

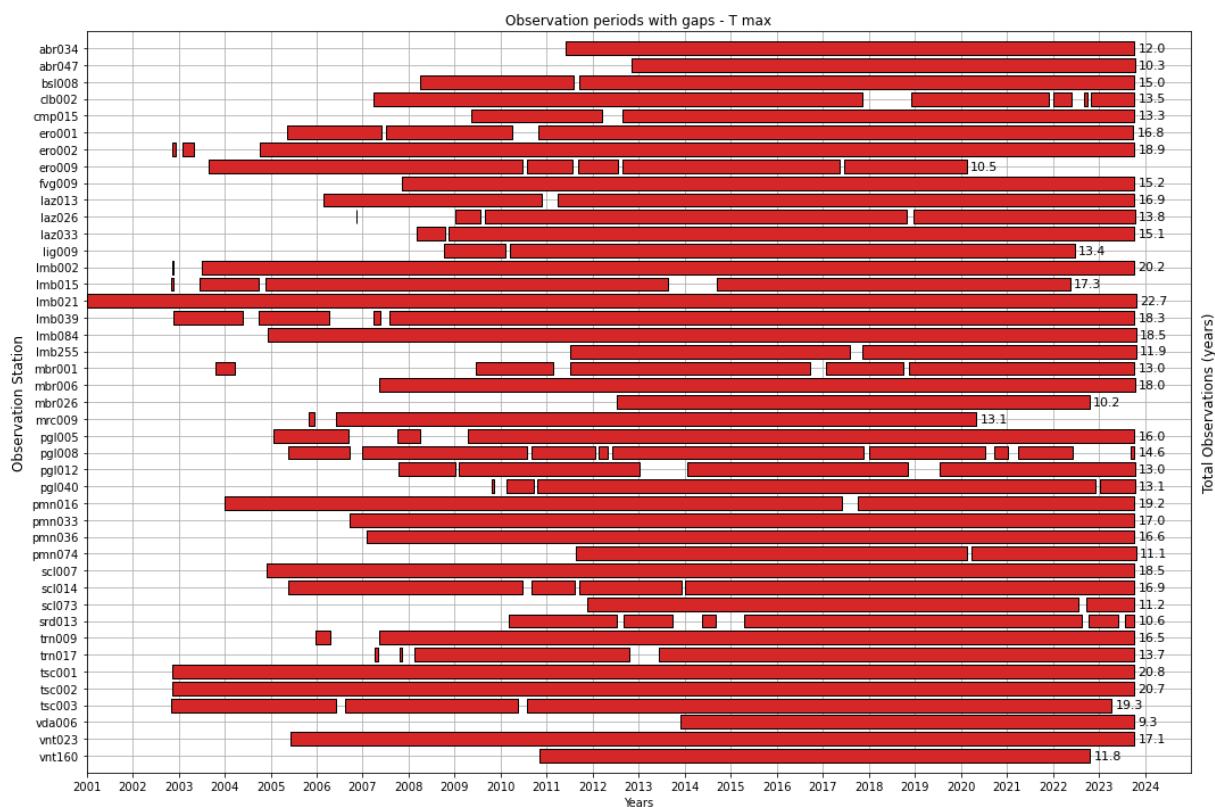


Figure 1: Observation periods of variable maximum temperature and effective length of historical series of each station in years.

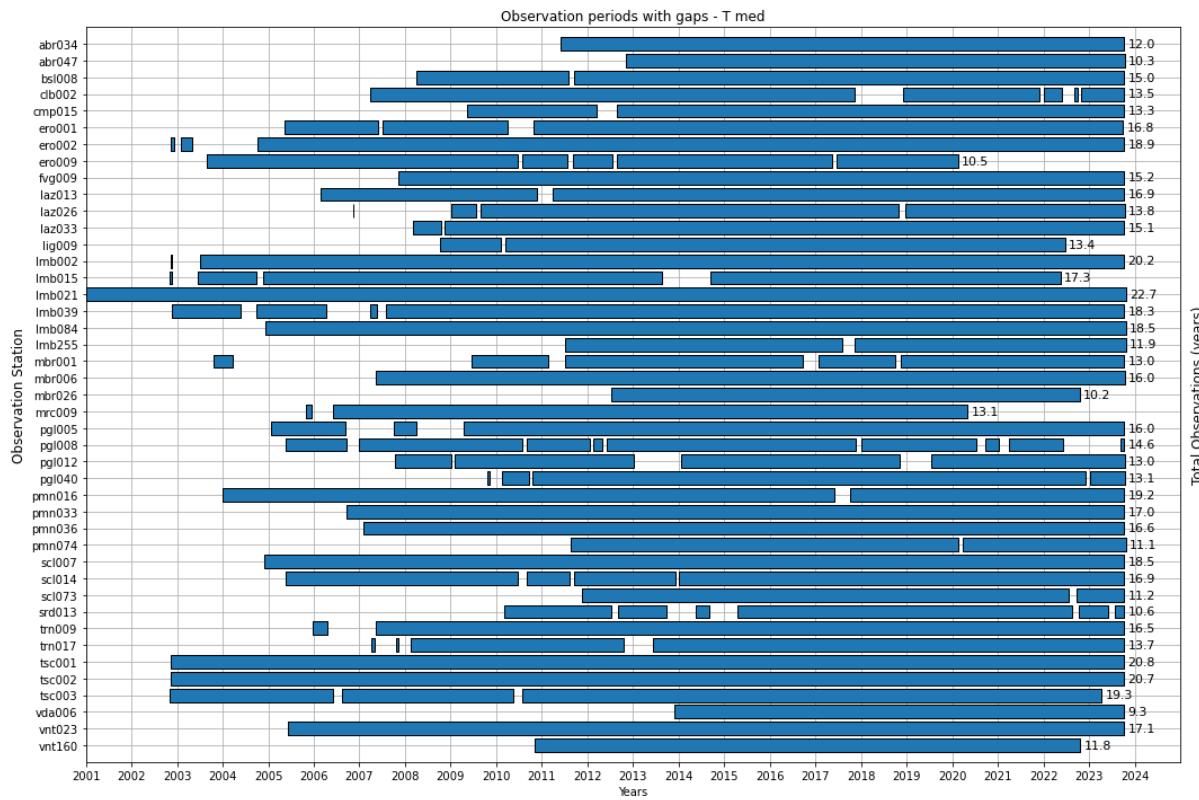


Figure 2: Observation periods of variable mean temperature and effective length of historical series of each station in years.

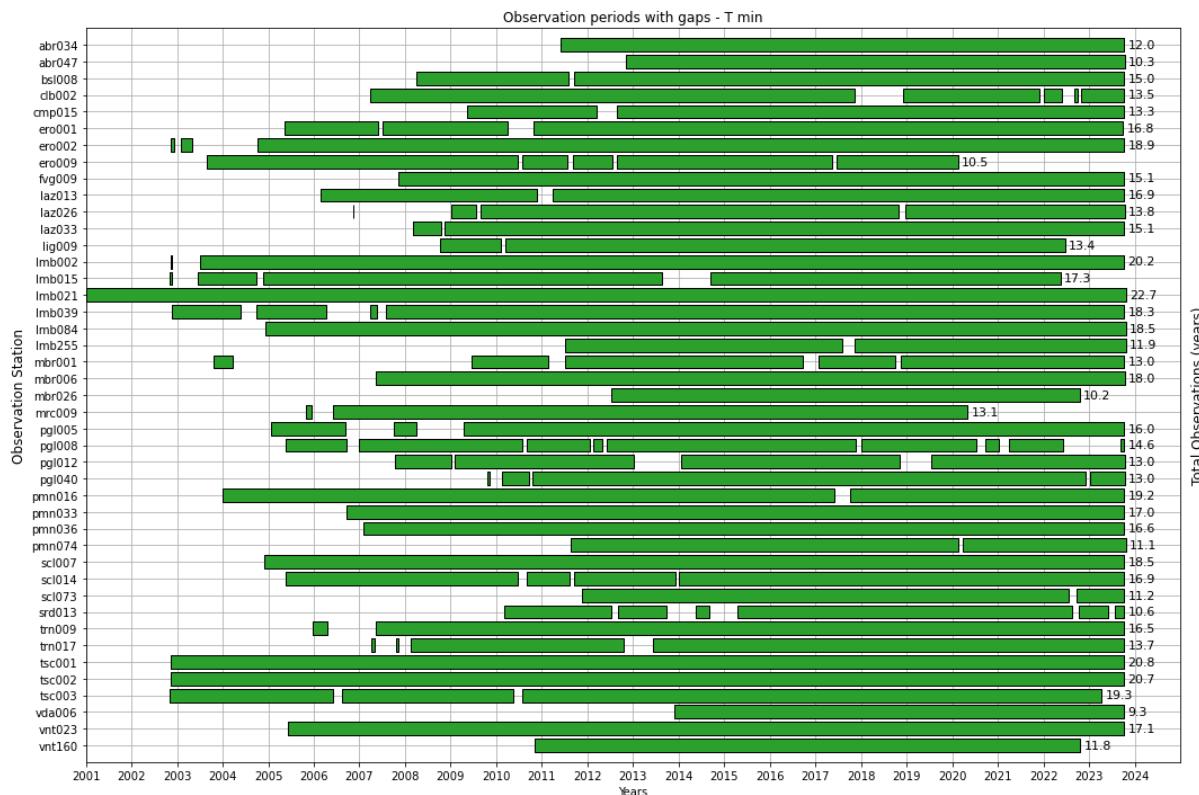


Figure 3: Observation periods of variable minimum temperature and effective length of historical series of each station in years.

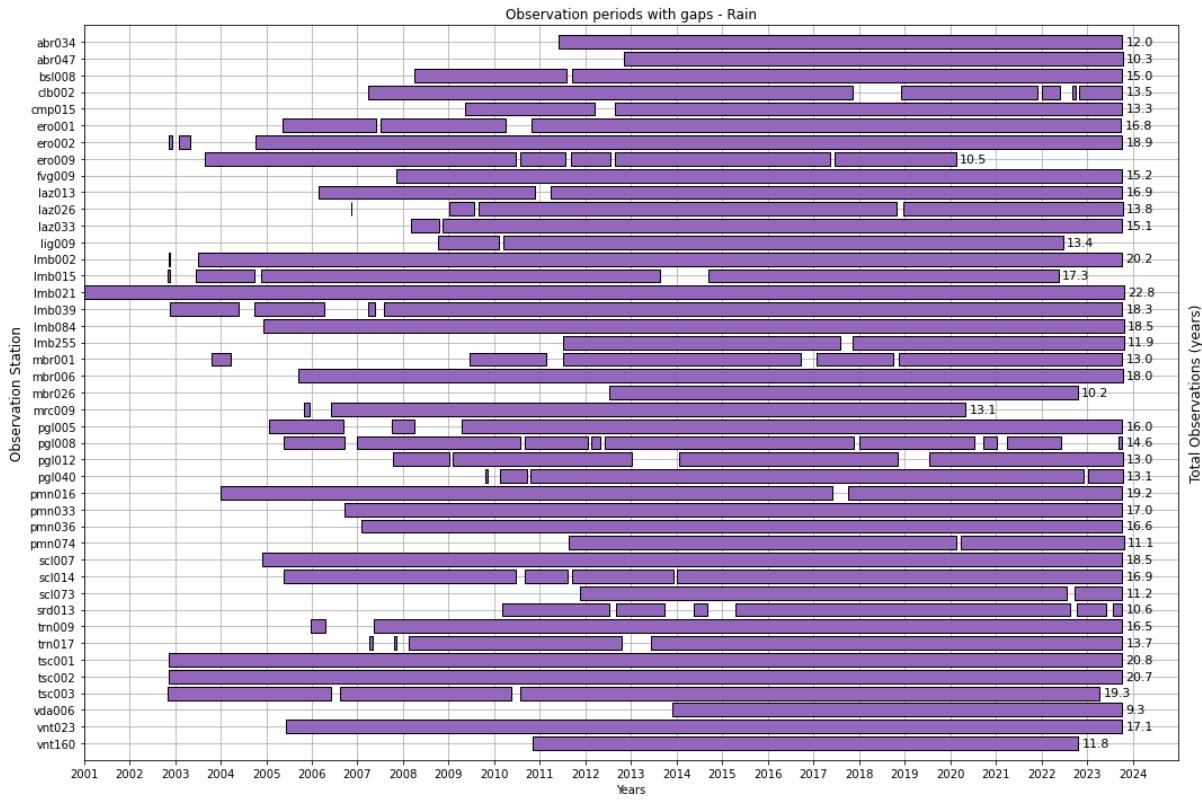


Figure 4: Observation periods of variable precipitation and effective length of historical series of each station in years.

Although there are some gaps in the series, the overall length of almost 10 years of observations for each variable is sufficient to enable the calibration of the Machine Learning model for each station to be continued with a reasonable degree of confidence.

### 2.2.2. Reanalysis Datasets

The categorisation of reanalysis datasets is typically based on the spatial domain. They can be classified into two principal categories:

- **Global reanalysis datasets**, which cover the entire surface of the planet;
- **Regional reanalysis datasets**, which are limited to specific geographic regions.

The selection of one of these two types of datasets is contingent upon the specific application. For instance, global datasets are frequently employed in the context of climate studies. In contrast, regional datasets are more frequently employed in national or subnational studies that necessitate greater specificity.

Global datasets are developed and provided by major meteorological organisations worldwide, which possess the requisite expertise and resources to manage and create highly complex modelling processes. One example of organisation engaged in the production of global reanalysis datasets is the European Centre for Medium-Range Weather Forecasts (ECMWF).

Global datasets provide data distributed over a large spatial grid, thereby making them optimal for the analysis of synoptic-scale climatic phenomena that affect large portions of the globe. Furthermore, these datasets can be employed as a foundation or boundary condition for regional-scale datasets.

Over the past few decades, numerous datasets have been developed and made accessible to users. The most notable of these is ERA5, which represents the benchmark for global reanalysis and also provides initial and boundary conditions for numerous regional reanalysis products. In 2019 ERA5 reanalysis data were released, under the auspices of the Copernicus Climate Change Service [21]. ERA5 is the fifth generation of ECMWF global reanalysis, covering the entire globe from 1979 at a spatial resolution of about 31 km.

In recent years, regional reanalysis datasets have assumed greater significance, driven by the growing demand for climate data tailored to specific areas or projects. The principal benefit of these datasets in comparison to global ones is their superior spatial and temporal resolution. Furthermore, they typically necessitate a considerably more compact grid, which considerably reduces the computational resources required for their development and increases the number of institutions or companies capable of producing them.

The principal characteristics that render reanalysis datasets so pervasive in contemporary usage are:

- The integration of a multitude of observational sources, if not all available ones, to create the datasets, with the objective of reducing errors to a minimum;

- A high degree of consistency with reality and historical measurements;
- Data coherence, ensured through various statistical corrections and cross-validation;
- Data provided on grids that are spatially and temporally uniform, regardless of the distribution and uniformity of the input data;
- Data is produced based on models that simulate atmospheric dynamics;
- Data allows for the retrieval of information on derived variables that are not directly measured but are obtainable through modelling;
- Data is easily accessible and freely usable by various users.

In the present study, five distinct reanalysis datasets were subjected to analysis with regard to temperature, comprising two global and three regional datasets. The global datasets are Era5-Land and MSWX, while the regional ones are CERRA, MERIDA H\_RES and VHR\_REA\_IT. In addition to the aforementioned reanalyses, two further datasets were considered for precipitation. The additional reanalysis datasets considered are MSWEP and CHIRPS.

#### 2.2.2.1. Era5-Land

In 2019, Copernicus Climate Change Service released ERA5-Land, produced by replaying the land component of the ECMWF ERA5 climate reanalysis. It implemented a series of improvements with respect to ERA5, in order to make it more accurate for all types of land applications. Indeed, ERA5-Land runs at enhanced spatial resolution respect to ERA5 (9 km vs 31 km). ERA5-Land uses as input to control the simulated land fields ERA5 atmospheric variables, such as air temperature and air humidity. Specifically, ERA5 air temperature, air humidity and pressure used to run ERA5-Land are corrected to account for the altitude difference between the grid of the forcing and the higher-resolution grid of ERA5-Land, according to the so called 'lapse rate correction' [21]. ERA5-Land covers the entire globe with hourly data from 1950 to about three months relatively to actual date. ERA5-Land data are produced with an

hourly time-step. Thus, daily values were derived by collecting 24 values starting from 00 UTC every day [21]. Some of the main products of this reanalysis are 2m temperature, evaporation, runoff, total precipitation, snow cover, soil temperature and many more. The products are available at the site: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form> (accessed on 1st September 2024).

ERA5-Land has the inconvenient that data, including atmospheric variables, are not provided for numerical grid points falling on sea surface. Thus, for land areas close to the coast, since they are not surrounded by four valid grid data values, the assessment of weather variables must rely on inland data only.

It should be also noted that the temperature is a genuine output of the ERA5-Land model, which may differ from the ERA5 dataset itself due to the potential influence of the land model. In contrast, surface precipitation is simply an interpolation of the ERA5 data on a more detailed grid, with no underlying physical or dynamic processes involved.

#### 2.2.2. CERRA

The Copernicus European Regional ReAnalysis (CERRA) is a regional reanalysis that has been produced for a domain covering entire Europe from 1984 to 2021. The reanalysis is produced as part of the Copernicus Climate Change Service. It is a high-resolution deterministic reanalysis characterised by a grid spacing of 5.5 km and that runs with a three-hour cycling. It uses a full set of in-situ observations and satellite information that adds quality while being reasonably consistent in time [22].

The inputs to CERRA reanalysis are the observational data, lateral boundary conditions from ERA5 global reanalysis as prior estimates of the atmospheric state and physiographic datasets describing the surface characteristics of the model [22].

Some of the main products available are 2m relative humidity, 2m temperature, albedo, evaporation, total precipitation and soil temperature. The data are available at

the site: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-cerra-singlelevels?tab=form> (accessed on 1st September 2024).

The added value of the CERRA data with respect to the global reanalysis products is expected to come, for example, with the higher horizontal resolution that permits the usage of a better description of the model topography.

#### 2.2.2.3. MERIDA HRES

MERIDA consists of a dynamic downscaling of the ECMWF global ERA5 reanalysis using the Weather Research and Forecasting (WRF) model, which is configured to describe the typical weather conditions of Italy. The ERA5 data are assimilated with a temporal resolution of 3 hours to ensure good temporal consistency. The temperature data from the SYNOP stations of the Italian Air Force are also assimilated into the WRF simulations with a temporal resolution of 3 hours. Furthermore, the optimal interpolation (OI) technique is applied to the modelled 2 m temperature and precipitation data through the use of meteorological observations of the Regional Agencies for Environmental Protection. The MERIDA domain consists of two grids with horizontal resolutions of 21 km and 7 km, respectively, with the inner grid centred on Italy [23]. MERIDA High-resolution for Renewable Energy Sources (HRES) is an upgrade of MERIDA more oriented towards describing the meteorological variables most commonly used in renewable energy applications. MERIDA HRES undergoes an increase in spatial resolution from 7 to 4 km. The products available are 2m temperature, total precipitation, wind components and solar radiation. The data are available at the site: [https://merida.rse-web.it/#download\\_new](https://merida.rse-web.it/#download_new) (accessed on 1st September 2024).

#### 2.2.2.4. MSWX

Multi-Source Weather (MSWX) is a seamless global gridded near-surface meteorological product with a high 3-hourly  $0.1^\circ$  resolution (about 9 km).

The product includes 10 meteorological variables: precipitation, air temperature, daily minimum and maximum air temperature, surface pressure, relative and specific humidity, wind speed, and downward shortwave and longwave radiation. The historical part of the record covers from 1 January 1979 to about 5 days from real time and is based on ERA5 data bias corrected and downscaled using high-resolution reference climatologies [24]. The products are available at the site: [www.gloh2o.org/mswx](http://www.gloh2o.org/mswx) (accessed on 1st September 2024).

#### 2.2.2.5. VHR\_REA\_IT

VHR\_REA\_IT (Very High Resolution REAnalysis for Italy) is dataset for recent climate developed within the Highlander project by dynamically downscaling ERA5 reanalysis to 2.2 km resolution (i.e., convection permitting scale). Dynamical downscaling was conducted through the COSMO Regional Climate Model (RCM). The temporal resolution of output is hourly (like for ERA5). Runs cover the whole Italian territory (and neighbouring areas according to the necessary computation boundary) to provide a very detailed (in terms of space–time resolution) and comprehensive (in terms of meteorological fields) dataset of climatological data for at least the last 40 years (01/1981-12/2020) [25]. Some of the main products available are surface evaporation, 2m temperature, total precipitation, wind components soil water contents and many others. The dataset is available at the site: [https://doi.org/10.25424/cmcc/era5-2km\\_italy](https://doi.org/10.25424/cmcc/era5-2km_italy) (accessed on 1st September 2024).

#### 2.2.2.6. MSWEP

Multi-Source Weighted-Ensemble Precipitation (MSWEP) is a recently released global *precipitation* dataset with a 3-hourly temporal resolution, covering the period 1979 to the near present. The dataset is unique in that it takes advantage of the complementary strengths of gauge, satellite, and reanalysis-based data to provide reliable  $P$  estimates over the entire globe. Since the release of version 1 ( $0.25^{\circ}$  spatial resolution) in May 2016, MSWEP has been successfully applied at global scales for a variety of purposes

[26]. MSWEP has several unique aspects: i) fully global coverage including all land and oceans (most satellite-based datasets are limited to 50° or 60° latitude); ii) high spatial (0.1°) and temporal (3 hourly) resolution, increasing the local relevance of the  $P$  estimates; iii) optimal merging of a wide range of gauge, satellite, and reanalysis  $P$  datasets, to obtain the best possible  $P$  estimates at any location; iv) correction for distributional biases, to eliminate spurious drizzle and restore attenuated peaks; v) correction of systematic terrestrial  $P$  biases due to gauge under catch using observed  $Q$  from 13,762 catchments worldwide; vi) corrections using daily observations from 76,747 gauges across the globe; and vii) a gauge correction scheme that accounts for gauge reporting times [26]. The data are available at the site: <https://www.gloh2o.org/mswep/> (accessed on 1st September 2024).

#### 2.2.2.7. CHIRPS

The Climate Hazards group Infrared Precipitation with Stations (CHIRPS) environmental record is a quasi-global (50°S-50°N), high resolution (0.05°), daily, pentadal, and monthly precipitation dataset. CHIRPS was developed to support the United States Agency for International Development Famine Early Warning Systems Network (FEWS NET). Building on approaches used in successful thermal infrared (TIR) precipitation products.

CHIRPS dataset builds on previous approaches to ‘smart’ interpolation techniques and high resolution, long period of record precipitation estimates based on infrared Cold Cloud Duration (CCD) observations. The algorithm i) is built around a 0.05° climatology that incorporates satellite information to represent sparsely gauged locations, ii) incorporates daily and monthly 1981-present 0.05° CCD-based precipitation estimates, iii) blends station data to produce a preliminary information product with a latency of about 2 days and a final product with an average latency of about 3 weeks, and iv) uses a novel blending procedure incorporating the spatial correlation structure of CCD-estimates to assign interpolation weights [27].

The data are available at the site: <https://earlywarning.usgs.gov/fews/datadownloads/Global/CHIRPS%202.0> (accessed on 1st September 2024).

### 2.2.3. Extraction of Reanalysis Data

In order to reconstruct the observation station series, it is necessary to provide the machine learning models with time series data from 1991 of daily maximum, average and minimum temperature and daily cumulative precipitation of the grid points of each reanalysis surrounding each station. All data series were extracted already aggregated on a daily scale from the aforementioned sites of each reanalysis.

The time series of observations from the grid points in closest proximity to each MNW station were extracted from all the aforementioned reanalyses. However, as the reanalyses have different resolutions and consequently different grids, to ensure that a similar number of points are extracted for all reanalyses, it was necessary to set a radius of 30 km for global reanalyses, while for regional reanalyses a radius of 20 km was set, so that the grid points extracted would be approximately 30 points for each reanalysis and station. The number of points extracted varied between 25 and 35, contingent on the location of the station and the location of the grid points of the reanalysis model.

In addition to differences in terms of spatial resolution, the reanalyses also exhibit variation in temporal resolution. Consequently, when the grid points were extracted (in October 2023), the longest available reanalysis series was selected for extraction. The aforementioned factors, namely the type of reanalysis, the duration required for processing data and any latency between the current date and the availability of data from the past, will influence the availability.

Consequently, the available data periods for each reanalysis are:

- MSWEP: from 1 January 1991 to 17 July 2023;

- CHIRPS: from 1 January 1991 to 30 June 2023;
- Era5-Land: from 1 January 1991 to 29 June 2023;
- MSWX: from 1 January 1991 to 29 December 2022;
- MERIDA\_HRES: from 1 January 1991 to 31 December 2021;
- CERRA: from 1 January 1991 to 29 June 2021;
- VHR\_REA\_IT: from 1 January 1991 to 31 December 2020.

It was decided to utilise as much of the available data from the reanalysis as possible in order to leverage the most recent data from the observation stations and to enable their use in the calibration of the model. Otherwise, the time series would have had to be reduced by more than three years in order to have data only up to 2020, resulting in the exclusion of pertinent data.

## 2.3. Machine Learning Models

### 2.3.1. Support Vector Regression

Support Vector Machine (SVM) is a supervised learning method used for both classification and regression. When used for regression, it is often called Support Vector Regression (SVR). SVR was initially proposed by Vapnik (1999), the objective of SVR is to find a function that approximates the input data with as little error as possible, using a minimum amount of error margin  $\epsilon$ . SVR attempts to find a line or a hyperplane, in the case of higher dimensions, that has at most  $\epsilon$  deviation from the actual data points. Assuming that the training dataset is given within the form  $[(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)]$ , where  $x_i$  presents the independent and  $y_i$  expresses the dependent variables, SVR transfers  $x$  to another feature space through a non-linear function  $\phi(x)$  and then approximates a function  $f(x) = w^T \phi(x) + b$  that not only has at most  $\epsilon$  deviation with  $y$  variable but also is as flat as possible [28],  $w$  and  $b$  are parameters of the

regression equation that need to be estimated to minimize the empirical risk function [19]  $R_{emp}$  as follows in Equation 1:

$$R_{emp} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_t|_\varepsilon \quad (1)$$

with:

$$|y_i - \hat{y}_t|_\varepsilon = \begin{cases} 0 & \text{for } |y_i - [w^T \cdot \phi(x_i) + b]| < \varepsilon \\ |y_i - [w^T \cdot \phi(x_i) + b]| - \varepsilon & \text{for } |y_i - [w^T \cdot \phi(x_i) + b]| \geq \varepsilon \end{cases}$$

The objective function is designed to minimise the prediction error while keeping model complexity low. This is done by minimising a cost function that considers errors above a certain margin  $\varepsilon$  and a penalty for model complexity. By minimizing the norm  $\|w\|^2$ , SVR can express the following convex optimization [29], shown in Equation 2:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T \cdot w + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (2)$$

with subject to:  $y_i - [w^T \cdot \phi(x_i) + b] \leq \varepsilon + \xi_i$ ,

$$[w^T \cdot \phi(x_i) + b] - y_i \leq \varepsilon + \xi_i^*$$

With:  $\xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, l$

where  $\xi_i$  and  $\xi_i^*$  are slack variables used to estimate the errors of the training samples outside error tolerance [19].

The main parameters characterising Support Vector Regression (SVR) are as follows:

- **C (penalty constant):** this parameter governs the trade-off between minimising training error and model complexity. A high value of C reduces the training

error, making the model less error tolerant, while a low value of  $C$  increases error tolerance, allowing the model to be more flexible;

- **$\varepsilon$  (insensitivity margin):** epsilon defines an insensitivity margin around the objective function, within which errors are not penalised. This parameter controls how much the model can ignore small errors;
- **Kernels:** kernels transform data into a higher-dimensional feature space, making the problem potentially linearly separable. Commonly used kernels include: Linear  $\langle x, x' \rangle$ , Polynomial  $(\gamma \langle x, x' \rangle + r)^d$ , Radial (RBF)  $\exp(-\gamma \|x, x'\|^2)$  and Sigmoidal  $\tanh(\gamma \langle x, x' \rangle + r)$ ;
- **Gamma ( $\gamma$ ):** this parameter determines the influence of a single training point on the model. High values of gamma imply a greater influence of the single point, while low values lead to a more diffuse influence.

Studies have demonstrated that the kernel function, regularization parameter  $C$ , and allowable tolerance of the epsilon tube are the main hyperparameters and are more sensitive to the final results [30],[31]. Accordingly, a series of experiments were conducted to ascertain the optimal values for these hyper-parameters for each station and variable.

### 2.3.2. Random Forest

The random forest algorithm, proposed by L. Breiman in 2001, has been extremely successful as a general-purpose classification and regression method [32]. Random Forest (RF) is a robust ensemble learning algorithm that uses bootstrapping to create multiple decision trees during training, introducing diversity either by sampling with replacement or by random feature selection, with the final output determined by aggregating the predictions of these individual trees. This approach reduces the risk of overfitting and improves the robustness of the model with respect to individual decision trees [32]. The Random Forest process is initiated with the acquisition of a dataset, which incorporates both the independent characteristics (features) and the

continuous dependent variable (target) that it is intended to predict. Subsequently, the construction of the decision trees is initiated through a process that initially involves bootstrap sampling. This method enables the generation of a multitude of data subsets through the extraction of samples with repetition, whereby some data may be selected on multiple occasions, while other data may be excluded.

A decision tree is constructed for each of the aforementioned subsets. The objective of each node within the tree is to identify the optimal split of the data, utilising a measure that aims to minimise variance. This may be achieved through the reduction of the mean square error, for instance. The process of node splitting is based on the selection of features, whereby a subset of the available features is randomly chosen for consideration, rather than all of them. The random selection is of great importance in ensuring the diversification of the constructed trees, as it serves to prevent them from becoming excessively similar to one another and to reduce the correlation between their predictions. If all trees were to utilise the same characteristics, they would be highly correlated and therefore less effective. Conversely, the diversity of trees permits a more robust and accurate final prediction. The subsequent phase is the aggregation of the results. Each tree provides an estimate of the continuous value of the target variable. The different estimates produced by the trees are then aggregated by calculating their arithmetic mean, which leads to the final prediction. Indeed, in regression problems, the RF algorithm returns predictions by calculating the mean of all Decision Trees [33].

The random selection of data and features also serves to enhance the model's capacity for effective generalisation to unobserved data. The construction of the trees in the Random Forest is independent of one another, which contributes to the robustness of the model. The number of trees included in the forest is a crucial element, as it is a hyper-parameter that must be carefully determined. In general, a greater number of trees tends to improve the performance of the model; however, beyond a certain point, a decrease in performance can be observed.

The main parameters characterising RF are as follows:

- **n\_estimators**: it indicates the number of trees in the forest;
- **max\_depth**: maximum depth of each tree. If not specified, trees are expanded until all leaves contain less than *min\_samples\_split* samples;
- **min\_samples\_split**: the minimum number of samples required to split an internal node;
- **min\_samples\_leaf**: the minimum number of samples required to be in a leaf node;
- **max\_features**: the number of features to consider when looking for the best division;
- **bootstrap**: if True, samples are sampled with replacement (bootstrap sample), otherwise all samples are used.

A series of experiments were conducted to ascertain the optimal values for the aforementioned hyper-parameters for each station and variable.

## 2.4. Performance Metrics

In order to investigate how similar, the different reanalysis datasets are to the observation datasets and to establish the efficacy of the proposed ML framework and the quality of the resulting temperature and precipitation estimates, an evaluation procedure has been implemented.

### 2.4.1. Taylor Diagram

The usual initial step in validating models is to determine whether their behaviour resembles the observed [34].

In order to select the most appropriate reanalysis dataset to be used as input in machine learning models, it was necessary to find a methodology for comparing the different datasets in an effective manner, taking into account a number of relevant

factors. The most straightforward and efficient method identified was the Taylor diagram. The Taylor diagram statistically quantifies the degree of similarity between two fields. One field will be called the "reference" field, usually representing some observed state. The other field will be referred to as a "test" field (typically a model-simulated field). The aim is to quantify how closely the test field resembles the reference field [34].

On this diagram the correlation coefficient and the root-mean-square (RMS) difference between the two fields, along with the ratio of the standard deviations of the two patterns, are all indicated by a single point on a two dimensional (2-D) plot [34]. It can provide a concise statistical summary of how well patterns match each other in terms of their correlation, their root-mean-square difference, and the ratio of their variances. Although the form of this diagram is general, it is especially useful in evaluating complex models, such as those used to study geophysical phenomena [34].

The statistic most often used to quantify pattern similarity is the correlation coefficient. Consider two variables  $x$  and  $y$ , which are defined at  $N$  discrete points, the correlation coefficient (CC) is defined as in Equation 3:

$$CC = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \quad (3)$$

where  $\bar{x}$  and  $\bar{y}$  are the mean values and  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ , respectively. The standard deviations are calculated as in Equation 4:

$$\sigma = \left[ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \right]^{\frac{1}{2}} \quad (4)$$

The correlation coefficient reaches a maximum value of 1 when for all  $n$ ,  $(x_n - \bar{x}) = \alpha(y_i - \bar{y})$ , where  $\alpha$  is a positive constant. In this case the two fields have the same centred pattern of variation but are not identical unless,  $\alpha = 1$ . Thus, from the correlation coefficient alone it is not possible to determine whether two patterns have the same amplitude of variation [34].

The statistic most often used to quantify differences in two fields is the RMSE, which for fields  $x$  and  $y$  is defined by Equation 5:

$$RMSE = \left[ \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \right]^{\frac{1}{2}} \quad (5)$$

A lower root-mean-square error (RMSE) indicates superior performance and lower error. The greater is the RMSE, the less are accuracy and precision of the estimates. The correlation coefficient and the RMSE provide complementary statistical information quantifying the correspondence between two patterns, but for a more complete characterization of the fields the variances (or standard deviations) of the fields must also be given [34]. With the above definitions it is now possible to construct a diagram that statistically quantifies the degree of similarity between two fields.

The diagram is constructed in a way that the radial distances from the origin to the points are proportional to the pattern standard deviations, and the azimuthal positions give the correlation coefficient between the two fields. The dashed lines measure the distance from the reference point and, indicate the RMS error [34]. Thus, as the model approaches the bottom right-hand side of the graph, it becomes increasingly in accordance with the observations. Additionally, as the model aligns with the curved line representing the standard deviation of the observations, the standard deviations of the two series exhibit a greater degree of similarity.

### 2.4.2. Bias

In addition, to compare the reanalyses and station observations, the bias was calculated, i.e. the average error of each reanalysis with respect to each station.

It was calculated as follows in Equation 6:

$$BIAS = \frac{1}{N} \sum_{i=1}^N (y_i - x_i) \quad (6)$$

where  $y_i$  is the model result and  $x_i$  is the observation, consequently, if the bias is positive, the model will overestimate the observations of that station; conversely, if the bias is negative, the model will underestimate the observations. The more the bias differs from zero the less accurate is the estimate.

### 2.4.3. Coefficient of Determination $R^2$

Two indicators were used to assess the performance of ML models. In addition to the previously illustrated RMSE, the coefficient of determination  $R^2$  was also calculated. The coefficient of determination is a statistical measure that indicates the goodness of fit of a model to observed data. The coefficient is calculated as one minus the ratio of the sum of the squares of the prediction errors (residuals), i.e. the difference between the observed values and the values predicted by the model, to the sum of the squares of the differences between the observed values and their mean, according to the following formula in Equation 7:

$$R^2 = 1 - \frac{\sum_{i=1}^N (x_i - \hat{y}_i)^2}{\sum_{i=1}^N (x_i - \bar{y})^2} \quad (7)$$

where  $x_i$  are the observed values,  $\hat{y}_i$  are the values predicted by the model and  $\bar{y}$  is the average of the observed values.

The value of  $R^2$  varies between 0 and 1. A value equal to 1 indicates that the model provides an optimal explanation of the observed variability in the data. A value equal to 0 indicates that the model is unable to explain the observed variability in the data, and that the use of the average of the observed values would have been an equally effective approach.

## 2.5. Methods

### 2.5.1. Comparison between Reanalysis Datasets and Observation Series

In order to obtain a model that most accurately reconstructs the time series of the MNW observation stations, it is fundamental to select a suitable input, consequently it is essential to select a reanalysis that demonstrates a high level of compatibility with the observations recorded at all stations, regardless of the length of the series or the specific characteristics of the station in question. To this end, each reanalysis dataset is compared with all observed data from all 43 stations collectively.

A merge must be performed between the reanalysis data of the previously extracted grid points and the observations of the stations, with the observation date serving as a common index. This is done to ensure that the resulting dataset exclusively comprises data from the specified common period. Indeed, this process excludes reanalysis data with dates that coincide with missing days in the station observations.

Global and regional reanalyses often show systematic, regionally distributed biases compared to ground-based weather station observations – especially of precipitation – which has led to the development of bias correction methodologies [35]. From a meteorological perspective, bias correction of reanalyses is widely considered an essential task to better reproduce spatial patterns of meteorological observations [12].

For this reason, in addition to comparing the raw reanalysis data without any modifications, simple procedures were applied to correct for possible biases that may occur, in order to understand which reanalysis is actually in closer agreement with the observations.

As a preliminary step, the bias was calculated for the temperature variables (T max, T med, T min) as previously described to determine the mean discrepancy between the values of the grid point in closest proximity to the station and the station's observations. Subsequently, the statistics of each reanalysis (CC, RMSE, and standard deviation) utilized in the Taylor diagram were calculated to construct the graphs of each variable for facilitating a visual comparison of the reanalyses with the observations. The initial comparison between the two datasets was conducted by considering solely the data from the grid point of the reanalysis in closest proximity to the station, with the objective of obtaining a 1:1 comparison between the model data and the observations. In order to select grid points according to their distance from the stations, an algorithm was devised to calculate the distances between points via their coordinates, with the objective of selecting the desired number of closest points from the total dataset.

It is typically insufficient to utilise a single grid point; therefore, the four closest points and then the sixteen closest points to the station were considered to gain more detailed information from the most proximate points. It is preferable to consider data from multiple sources, as it is uncertain whether the nearest grid point represents the location with the most accurate observations, even though the high density of the reanalysis grids and the proximity of the points, which are no more than a few kilometres apart. This phenomenon is exemplified in instances where the station and grid point are in close proximity but situated at disparate altitudes. This can occur in mountainous regions, where abrupt changes in altitude may occur over short distances. The case in question presents certain difficulties with regard to temperature, e.g. when a meteorological station is situated in a valley and the grid point is located

on a mountain slope, there is the potential for a significant temperature differential to occur. Therefore, it is preferable to utilise data from multiple grid points, when available, to limit this particular case.

In order to aggregate the values from several points in a way that allows for a 1:1 comparison and ensures that the series to be compared are of the same length, two different methods were employed for aggregating the grid point values: a simple arithmetic mean of the values of the four and sixteen points was employed to assign equal weight to all grid points, while Inverse Distance Weighting (IDW), an interpolation method, was used to account for the distance of the reanalysis points from the station, thereby assigning greater weight to the data from the closest points. To place special emphasis on distance, the weights were assigned considering the inverse of the square of the distance.

The following formula in Equation 8 was used to apply IDW:

$$IDW = \frac{\sum_{i=1}^n \left( \frac{z_i}{d_i^2} \right)}{\sum_{i=1}^n \left( \frac{1}{d_i^2} \right)} \quad (8)$$

where  $z_i$  is the observation value of the reanalysis grid point and  $d_i$  is the distance between the station and the grid point.

The methodology previously outlined for the retrieval of the series utilized in the calculation of the statistical values represented in the Taylor diagram was employed for both temperature and precipitation variables.

Furthermore, an altitude correction was applied to the values of the grid points for the temperature variables only, prior to the application of the IDW, ensuring a better comparison between reanalysis and observations. As the grid points of the reanalysis are not situated at the same altitude as the observation station, it is necessary to correct

the temperature values of the grid points in order to account for the dependence of temperature on altitude. In order to correct the temperature for the difference in elevation, it is first necessary to determine the temperature lapse rate, which indicates how the temperature varies with elevation. The elevation differential between the station and each of the sixteen selected grid points is calculated. The lapse rate is determined separately for each station and for each date, as it can vary considerably depending on local atmospheric conditions, such as humidity, air stability and the presence of thermal inversions. Consequently, in specific studies or in particular regions, it may be necessary to employ a lapse rate that has been locally observed or calculated. In the literature, the lapse rate is typically represented by an average value of 6.5 °C per kilometre (or 0.65 °C per 100 metres). This value, designated the standard atmospheric lapse rate, is employed in numerous meteorological and climatic applications to represent the mean temperature decline with altitude in a standard atmosphere. However, in this study, it was decided that a locally calculated value would be more accurate.

The lapse rate for each station is calculated through a linear regression, performed separately for each date. For each date, the 16 reanalysis points and their temperature values are considered. The altitude of the grid point is the independent variable ( $X$ ), while the temperature simulated by the model is the dependent variable ( $y$ ). A linear regression model is fitted to the data in order to estimate the lapse rate, which is represented by the coefficient of the  $X$  variable, represented in the Equation 9 by the coefficient  $m$ .

$$y_i = mX_i + q \quad (9)$$

Once the lapse rate has been determined, it is employed to correct the temperature simulated by the model, taking into account the altitude difference. The temperature change ( $\Delta T$ ) is calculated by multiplying the lapse rate by the altitude difference ( $\Delta H$ ), shown in Equation 10:

$$\Delta T = m * \Delta H \quad (10)$$

In the event that the station is situated at a greater altitude than the grid point, the temperature is reduced by subtracting  $\Delta T$ ; conversely, if the station is located at a lower altitude, the temperature is increased by adding  $\Delta T$ , as shown in Equation 11:

$$y_{corr} = y \pm \Delta T \quad (11)$$

Subsequently, a weighted average of the corrected temperatures is calculated using IDW, and is then compared with the actual station observations. The resulting statistics are presented in Taylor diagrams.

The altitude correction was only applied to the Era5-Land, CERRA and MERIDA H\_RES reanalyses, as these were the only reanalyses for which the altitude data for each grid point could be found.

In the next chapter the results obtained will be presented: the bias maps, the Taylor diagrams and the choice of which reanalysis to use as input in the ML procedure.

### 2.5.2. Machine Learning Model Optimization

Once the reanalyses for both temperature and precipitation had been identified as input for the ML models, the subsequent step was to design the model, selecting the

optimal model type for each variable and optimising the parameters for each specific station.

It was resolved that a single model should be constructed for each of the 43 stations and for each of the four selected variables (T max, T avg, T min, P). This resulted in the creation of 172 models, which were developed with the objective of enhancing the accuracy of the models and, subsequently, the climatologies. The specificity of each model to a given station and variable is made possible by the relatively short training time of the ML model, which allows for the repetition of the training procedure an adequate number of times. If the model in question had been a classical physically-based one, this would not have been feasible.

A distinctive feature of ML models is their ability to identify which inputs have the greatest influence on the prediction of output values. For this reason, all available reanalysis grid points around each station were used as inputs, allowing the model to determine which of these inputs most significantly affect the outputs.

The two distinct ML models outlined previously, SVR and RF were subjected to evaluation. The models were tested by varying the values of the hyper-parameters outlined in the previous chapter.

Overfitting in a model generally provides high accuracy on the training dataset; however, it displays poor performance on other datasets. To deal with this problem, hyperparameters in each ML algorithm should be tuned to obtain the models, which could be generalized data in both the training and validation datasets. The suitability hyper-parameters for the models were verified through the goodness of fit of the results on the validation data based on statistical measures in terms of accuracy through the coefficient of determination  $R^2$  and Root Mean Squared Error (RMSE).

Each model was trained on the common period of each station and the chosen reanalysis. In order to train the model, the data were set as follows: the independent variables (X) consist of the values at different observation points of the reanalysis,

while the dependent variable ( $Y$ ) is the observed value of the station. The sample data is given by all daily observations.

A dataset was created for each station variable, comprising the inputs (reanalysis data) and outputs (observations). This was randomly divided into two parts: 80 per cent was used for calibration and 20 per cent for validation.

In the initial stage of the experiment, a particular type of ML model is selected for testing purposes. An initial plausible value is then assigned to the parameters in order to verify the performance of the model. Subsequently, a number of attempts are made using different parameter values identified in the literature. A Grid Search approach was employed in order to enable an exhaustive search for the optimal hyperparameters.

Grid search is a technique employed for the purpose of optimising the parameters of a machine learning model. The technique entails the testing of a combination of disparate parameter values, the subsequent evaluation of the model's performance with each combination, and the selection of the combination that provides the optimal results based on a specific evaluation metric. The evaluation metric employed is the optimal  $R^2$ .

The process of how Grid Search works is the following:

1. A grid of values is created for each parameter, representing a range of plausible values that the parameter can assume. Each potential combination of these values will be evaluated;
2. The model is evaluated for each parameter combination using the technique of cross-validation. This process helps to ensure that the results are robust and not overly dependent on the specific training data set;

3. The selection of the optimal parameters is a crucial step in the process. The combination of parameters that exhibits the optimal average performance on cross-validation is selected as the optimal choice;
4. The final model is then trained, using all the training data with the optimised parameters.

Cross-validation is a verification technique that assesses the model's capacity for generalisation when applied to an independent dataset. It delineates a dataset that is employed for the purpose of testing the trained model during the training phase in order to prevent the phenomenon of overfitting. In this study, the k-fold cross-validation technique was employed. The training dataset was partitioned into k mutually exclusive subsamples (or folds) of equal size. The model was trained k times, with each iteration using one of the k subsamples for testing, and the remaining k-1 subsamples applied towards training the model. The k results of cross-validation are averaged to estimate the accuracy as a single estimation. The value of k was set equal to five.

To identify the optimal hyper-parameter values, an iterative process was employed, commencing with broad ranges encompassing typical literature parameter values. This was undertaken to initially delineate an approximate zone of parameter values that would yield satisfactory performance. Subsequently, the grid was narrowed around the identified best values, and this process was continued until the  $R^2$  value demonstrated a plateau and remained relatively stable.

The optimal cross-validation score on the validation data is determined by calculating the mean of the  $R^2$  values across the five distinct folds for the parameter set that demonstrated the most favourable overall performance. This metric was employed for the purpose of selecting the optimal model.

The following procedure was followed for each of the three temperature variables ( $T_{avg}$ ,  $T_{max}$ ,  $T_{min}$ ) and for each station individually. A Grid Search approach was

initially employed for the SVR model, enabling an exhaustive search for the optimal hyperparameters, including the penalty parameter C, epsilon term, gamma, and linear kernel or RBF. The search is guided by an evaluation metric based on R<sup>2</sup> and root mean squared error (RMSE), with the optimal model refit based on R<sup>2</sup>. The performance of the model was evaluated on the training and validation data sets by calculating the root mean squared error (RMSE) and the mean R<sup>2</sup>. The optimal parameters and the optimal model are stored for future use.

The following example illustrates the structure of a value grid for SVR model parameters:

$$\text{param\_grid} = \left\{ \begin{array}{l} C: [1, 10], \\ \gamma: [1E - 6, 1E - 1], \\ \text{kernel: [linear, rbf]}, \\ \varepsilon: [1E - 3, 1E - 1] \end{array} \right\}$$

Similarly, the Random Forest model is trained and optimised using a Grid Search approach, which is analogous to the SVR approach. The hyperparameters that were considered included the number of trees (*n\_estimators*), the maximum depth of trees (*max\_depth*), the minimum number of samples required to split a node (*min\_samples\_split*), the minimum number of samples required for a leaf node (*min\_samples\_leaf*), and the use of bootstrapping. Once more, the optimisation process is driven by the R<sup>2</sup> and RMSE metrics. The performance metrics, including the root mean squared error (RMSE) and the mean R<sup>2</sup>, are calculated and displayed for both the training and validation phases. In conclusion, the optimal Random Forest model is stored for subsequent use.

The following example illustrates the structure of a value grid for RF model parameters:

$$\text{param\_grid} = \left\{ \begin{array}{l} n\_estimators: [100, 200], \\ max\_depth: [10, 20], \\ min\_samples\_split: [5, 10], \\ min\_samples\_leaf: [2, 4], \\ bootstrap: [True, False] \end{array} \right\}$$

### 2.5.2.1. Adding Sample Weights to Precipitation Models

Precipitation is a more complex phenomenon to model than temperature due to the intrinsic nature of the physical processes involved. In most cases, precipitation does not occur, resulting in recorded data that is typically zero. This leads to a pronounced skewness in the distribution of possible values, which can challenge the ability of the model to accurately represent such trends.

To address this issue, it was determined that the use of sample weights would be an effective approach for training a model that is specifically focused on precipitation days, while excluding days with zero precipitation.

The weight assigned to days with precipitation is significantly higher than that assigned to days without precipitation, with the latter receiving a weight of zero.

During the training phase, each data point is multiplied by its respective weight. Data points with zero weight do not contribute to the model error and therefore have no impact on the model parameters. Conversely, data points with high weight contribute significantly to the error and therefore exert a considerable influence on the training process. This approach is beneficial for enhancing the model's capacity to forecast precipitation.

By focusing on days with precipitation, the model is better able to discern the patterns and relationships that lead to precipitation. Indeed, in many meteorological datasets, the number of days without precipitation may far exceed the number of days with precipitation. This imbalance may result in the model exhibiting a tendency to neglect precipitation days. The application of weights serves to compensate this imbalance. Furthermore, the absence of precipitation may introduce noise into the model, thereby impeding its capacity to discern patterns pertinent to precipitation. By disregarding these days, the model can concentrate solely on the pertinent data.

Through a process of trial and error, it was decided to assign the weight of  $10^4$  to days with precipitation.

Once precipitation has been assigned a weight, the procedure followed for rainfall is equivalent to that used for temperature. The model's performance is evaluated on both the training and validation datasets, with the average root mean squared error (RMSE) calculated on the five folds into which the datasets are divided. Similarly, the mean R<sup>2</sup> was determined for the aforementioned datasets. In conclusion, the optimal model, based on the most suitable hyperparameters identified, is saved for future use.

### 2.5.3. Model Application and Reconstruction of Climatologies

Once the optimal model for each station's variable has been found, it can be employed to generate a reconstructed time series for the variable of interest at each station, for all periods where reanalysis data is available. Reanalysis data were extracted for the reference period 1991-2020. These data were used as input for the selected machine learning model, which had already been calibrated. The optimal model found for each specific variable (mean temperature, maximum temperature, minimum temperature, or precipitation) was then employed to generate the predictions. This model, trained with reanalysis data, reconstructs the time series for missing or incomplete weather station observations for all periods between 1991 and 2020. Once the complete series was reconstructed, the following calculations were performed:

- **Monthly climatologies for temperature variables:** the daily data were then grouped by month, and the monthly average for each temperature variable was calculated, resulting in a monthly climatology. This represents the average mean, maximum, and minimum temperature for each month over the period 1991-2020;
- **Monthly climatologies for precipitation:** the sum of daily precipitation was calculated for each day of the year. The daily precipitation data were then aggregated on a monthly and yearly basis to calculate the monthly cumulative precipitation totals. Finally, the mean monthly cumulative precipitation for the

entire period was calculated, representing the average total precipitation for each month.



# 3 Results and Discussion

## 3.1. Comparison of Reanalysis Datasets

### 3.1.1. Temperature Bias

The analysis of the biases of temperature variables calculated using only the nearest grid point of each reanalysis for each station reveals several key findings. In general, all reanalyses tend to underestimate maximum temperatures, with the exception of VHR, which tends to overestimate them. The stations exhibiting the most pronounced biases in maximum temperatures are situated in the elevated regions of the Alps. It is noteworthy that the station *trn017* in Trentino Alto Adige, situated at an altitude of over 2,000 metres, consistently exhibits one of the highest overestimations across all reanalyses, with an average difference of at least 4°C for maximum temperature. However, CERRA exhibits a slightly lower bias of approximately 3°C. The station exhibiting the greatest bias in absolute terms is *lmb255* in Lombardy, situated at over 1,300 meters, where maximum temperatures are underestimated by an average of 8°C across all reanalyses. With respect to maximum temperatures, Era5-Land demonstrates the most substantial divergence from observations when all stations are taken into account.

With regard to average temperatures (Figures 6-8), the most considerable biases are also observed in the mountainous Alps region, with Era5-Land continuing to demonstrate the highest average biases. The station *lmb255* continues to exhibit the greatest underestimation of average temperatures, with a bias of approximately 7°C, while *trn017* displays the most significant overestimation, with a bias of around 3°C.

However, the biases for average temperatures are, in general, smaller than those for maximum temperatures.

In the case of minimum temperatures, the VHR data set exhibits the greatest discrepancy from the observed values. The bias observed in the reanalysis of station *lmb255* is approximately 6.5°C, whereas the bias in the reanalysis of *trn017* is considerably smaller, averaging approximately 1.5°C. The station exhibiting the greatest degree of overestimation varies according to the reanalysis employed, but there is a discernible tendency for overestimation in eastern Italy. Overall, the largest biases are observed for maximum temperature, followed by minimum temperature, with the smallest biases found in average temperature. Ultimately, altitude represents the most significant factor influencing discrepancies between reanalysis data and observations, while proximity to the sea has a negligible impact.

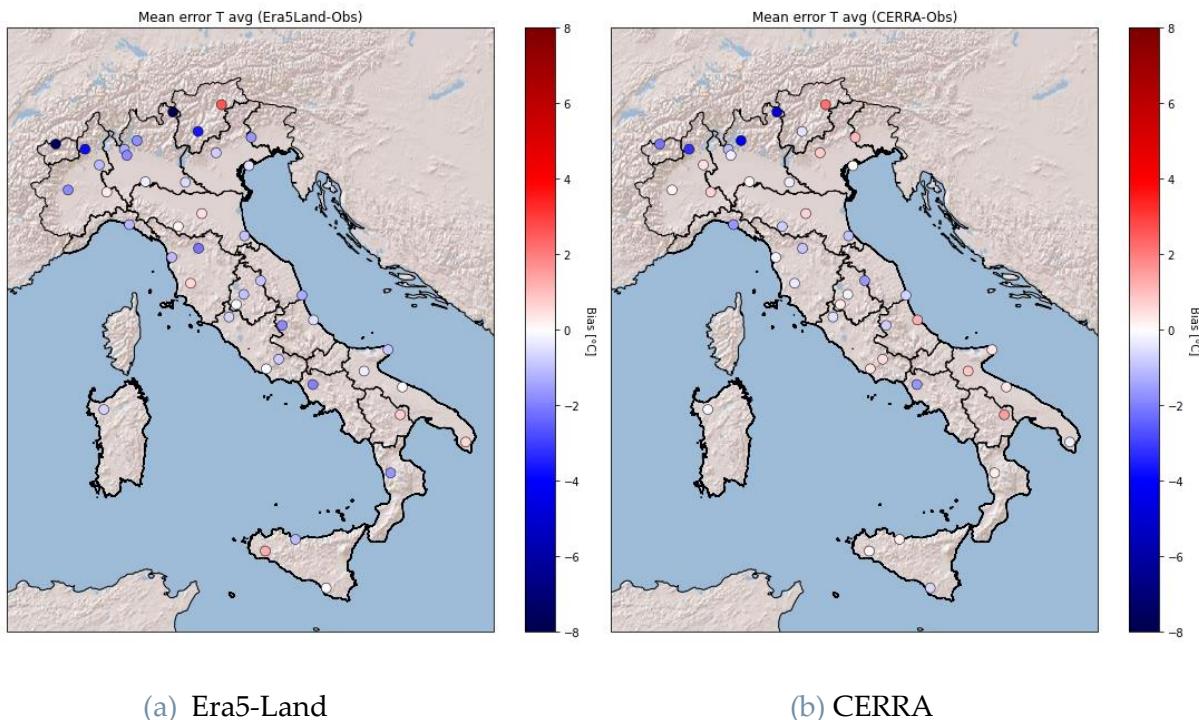
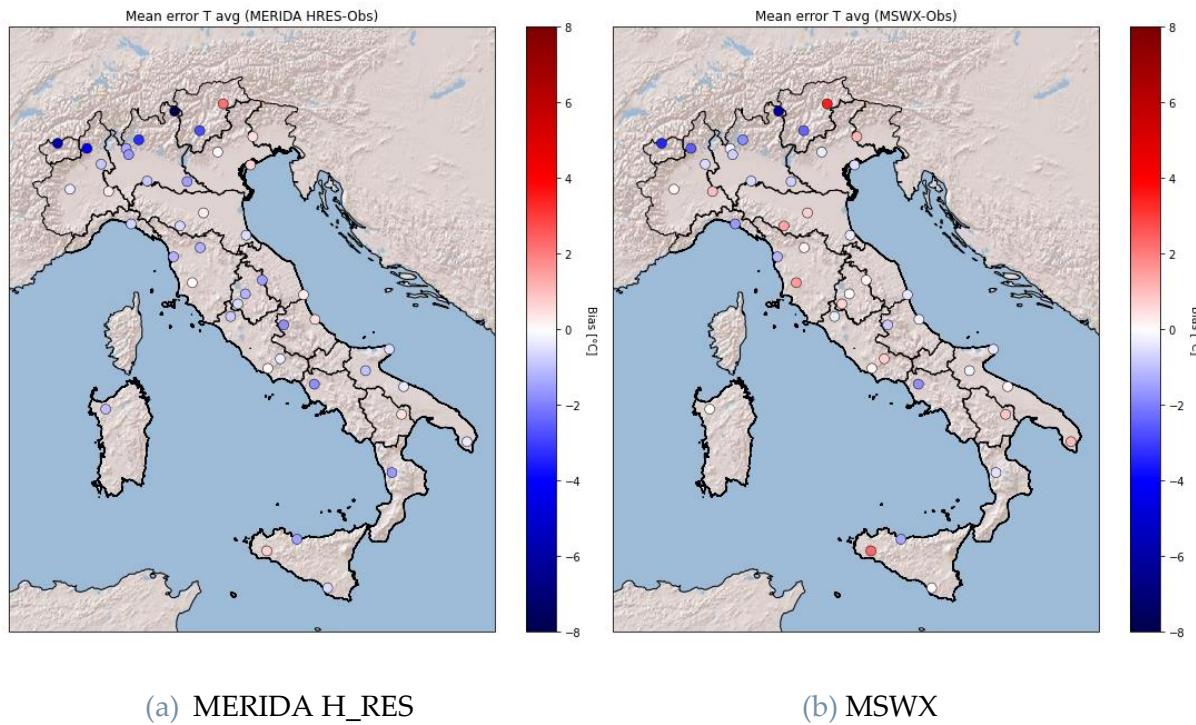


Figure 6: Mean biases of the average temperature of each station of Era5-Land and CERRA compared to the observations, considering only the nearest grid point.



(a) MERIDA H\_RES

(b) MSWX

Figure 7: Mean biases of the average temperature of each station of MERIDA H\_RES and MSWX compared to the observations, considering only the nearest grid point.

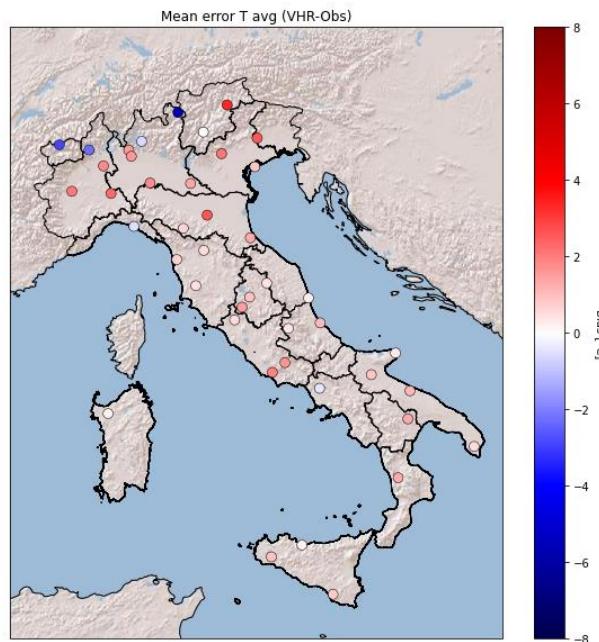


Figure 8: Average bias of the average temperature of each station of VHR compared to the observations, considering only the nearest grid point.

### 3.1.2. Taylor Diagrams

The Taylor diagrams obtained considering only the nearest grid point are shown in Figure 9. There, it can be seen that the results of the statistics are promising and indicate that there is a high level of agreement between the reanalyses and the observations.

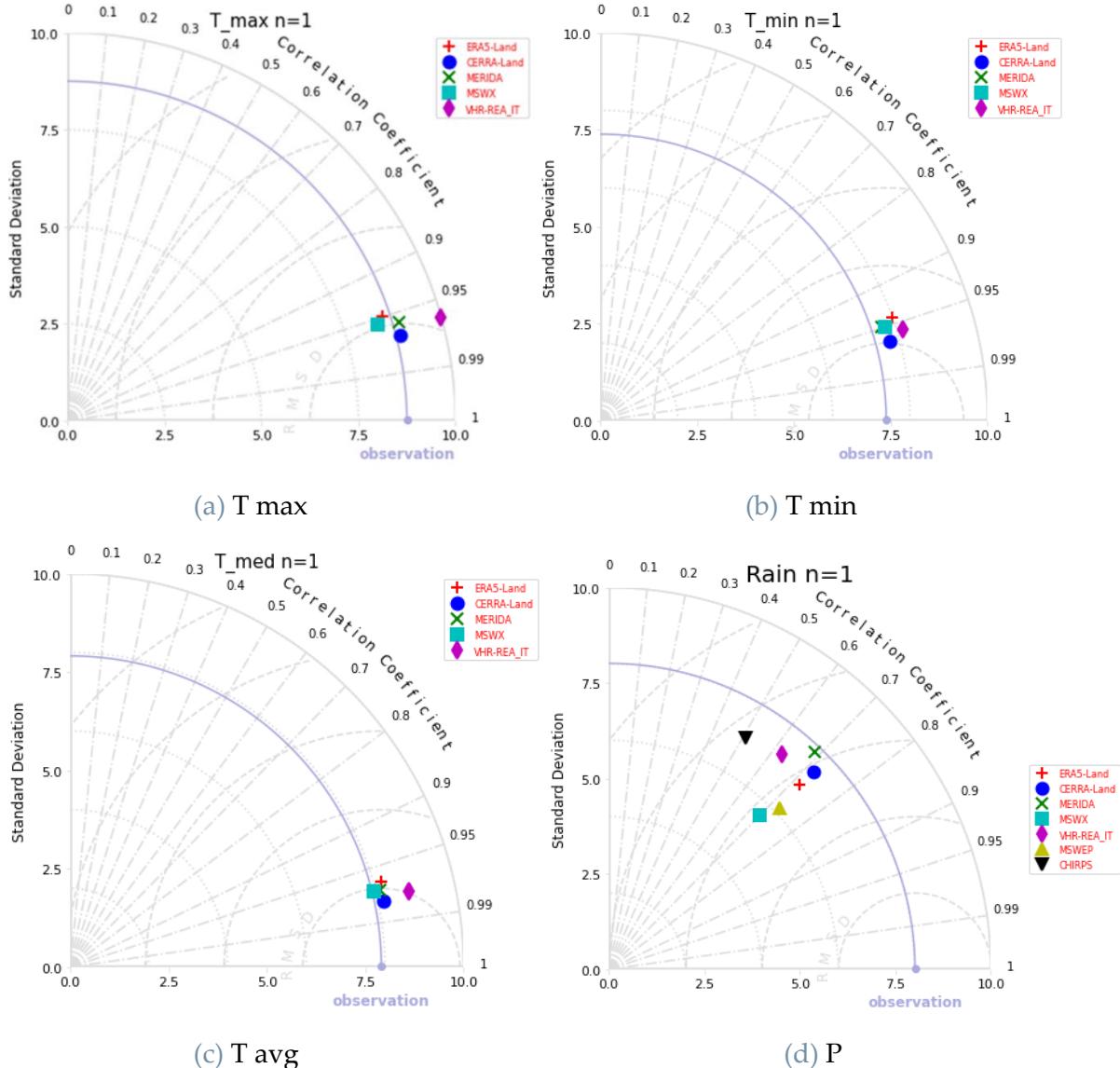


Figure 9: Taylor diagrams of the variables T max, T avg, T min and P of each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest grid point.

For the temperature variables, the correlation coefficient is consistently above 0.9 for all reanalyses, while the root mean square error (RMSE) is invariably below 3°C. The

CERRA dataset exhibits the most favourable correlation coefficient and RMSE values for all three temperature variables.

With regard to standard deviations, the MSWX reanalysis dataset has the closest value to the observed standard deviation for the average and minimum temperature, while for the maximum temperature, the CERRA dataset displays the most similar value. The degree of similarity between the standard deviations of the model and the observations visibly varies depending on the reanalysis and on the variable.

The variable most closely aligned with the reanalysis is, not surprisingly, the average temperature, followed by the maximum and minimum temperature. It can be observed that all the reanalyses demonstrate a high degree of compatibility, regardless of their spatial or temporal resolution. Indeed, with regard to temperature variables, the reanalyses exhibit greater similarity and less deviation. VHR is the reanalysis that appears to exhibit the poorest performance, with the greatest discrepancy from observed values, particularly for maximum temperatures.

With regard to precipitation, there is a considerable discrepancy in the correlation coefficient (CC), with values ranging from 0.5 for CHIRPS to 0.73 for MSWEP. Additionally, the RMSE is between 5.5 mm (MSWEP) and 7.5 mm (CHIRPS), while the standard deviations exhibit a notable variation depending on the reanalysis. In fact, the reanalyses display a considerably greater degree of scatter. The MERIDA H\_RES reanalysis dataset demonstrates the closest approximation to the standard deviation of the observations. The reanalysis that deviates the most from the observations is CHIRPS, followed by VHR.

The reanalyses that are most consistent with the observations are CERRA, MERIDA H\_RES, ERA5-Land and MSWEP. In particular, CERRA is the superior reanalysis in terms of correlation coefficient and root mean square error (RMSE) for all temperature variables, whereas MSWEP is the more accurate reanalysis for precipitation.

As illustrated in Tables 4 and 5 and in Figure 10, an increase in the number of grid points considered has a positive impact on performance in terms of CC and RMSE for precipitation, but not for temperature. The two indicators demonstrate an increase for precipitation and a slight decrease for temperature. Conversely, the standard deviations of the series demonstrate an inverse relationship, exhibiting a tendency to converge as the number of grid points increases for temperature while diverging for precipitation.

Table 4: Taylor diagram statistics for the CERRA reanalysis of the variable T avg, considering 1 and 16 points.

T avg	N=1		N=16	
	Observation	Model	Observation	Model
CC	1	0.978	1	0.974
RMSE	0	1.687	0	1.836
Std. Dev	7.869	8.144	7.869	8.131

Table 5: Taylor diagram statistics for the MSWEP reanalysis of the variable P, considering 1 and 16 points.

P	N=1		N=16	
	Observation	Model	Observation	Model
CC	1	0.729	1	0.743
RMSE	0	5.497	0	5.368
Std. Dev	8.018	6.152	8.018	5.991

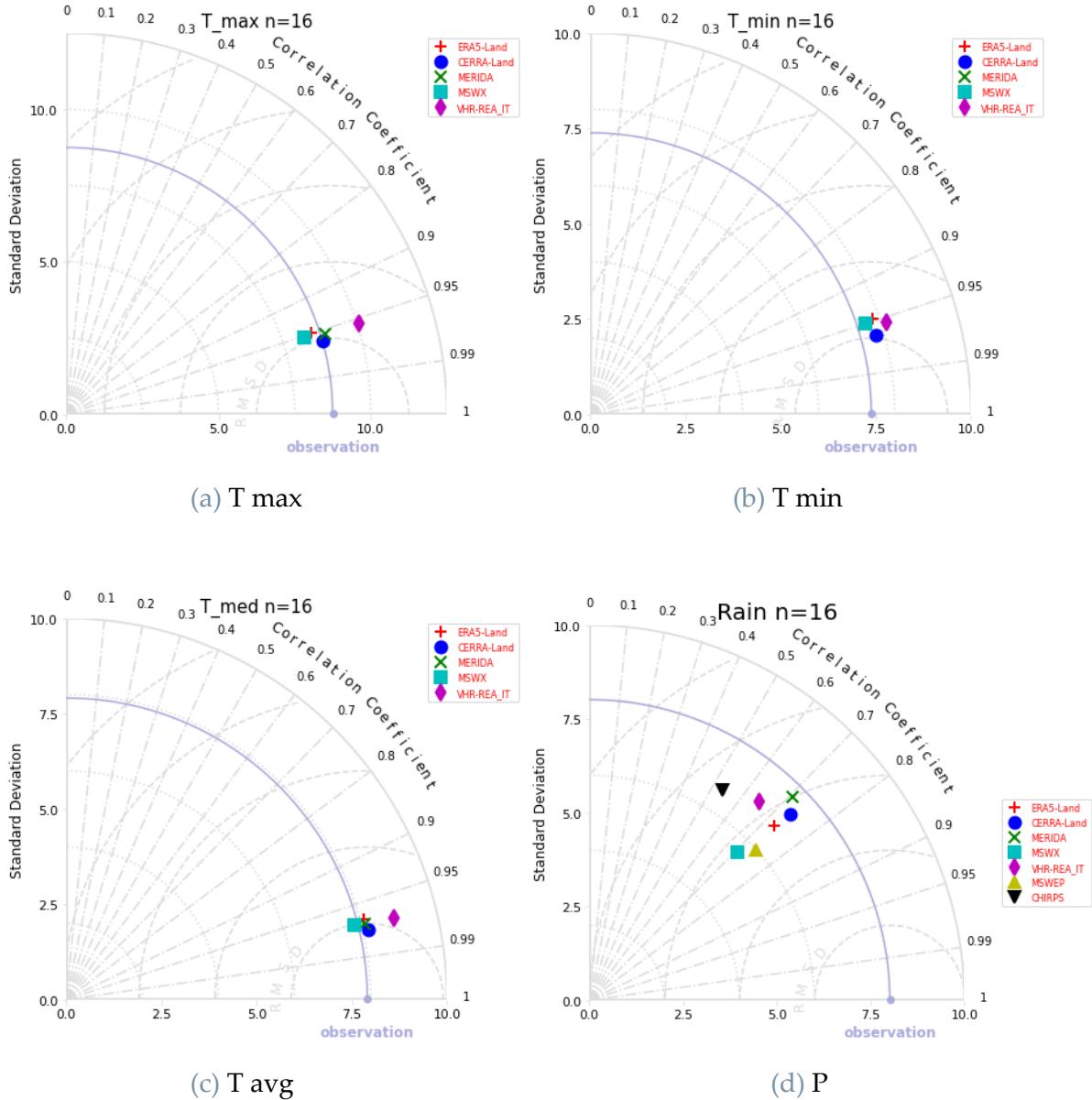


Figure 10: Taylor diagrams of the variables T max, T avg, T min and P of each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest 16 grid points, aggregated with arithmetic mean.

The values displayed in Table 5 represent the highest degree of agreement between the reanalysis models and the observations for the precipitation variable in the case where N is equal to 16, as indicated by the correlation coefficient and the root mean square error (RMSE).

On the other hand, when considering the distances between the points, as illustrated in Figure 11 and presented in Table 6, it becomes evident that for temperature,

distances enhance the agreement between reanalysis and observations, improving both the correlation coefficient and the RMSE. At the same time, in contrast to the previous case, this results in a reduction in the standard deviation. This allows for better performance when distances are considered via the inverse distance weighting (IDW) method. A comparison of the case with four points and IDW with the case with 16 points and IDW reveals that the two cases exhibit minimal differences. The correlation coefficients are identical, the root mean square error (RMSE) with four points is slightly lower, even though the standard deviations are more similar in the case with 16 points.

Table 6: Taylor diagram statistics for the CERRA reanalysis of the variable T avg, considering 16 points aggregated with arithmetic mean and with IDW.

T avg	N=16		N=16 IDW	
	Observation	Model	Observation	Model
CC	1	0.974	1	0.979
RMSE	0	1.836	0	1.675
Std. Dev	7.869	8.131	7.869	8.118

In contrast as shown in Figure 11 and Table 7, the incorporation of distance data into precipitation calculations results in a convergence of the standard deviations, although this approach also yields a slight decline in the correlation coefficient and RMSE performance.

Table 7: Taylor diagram statistics for the MSWEP reanalysis of the variable P, considering sixteen points aggregated with arithmetic mean and with IDW.

P	N=16		N=16 IDW	
	Observation	Model	Observation	Model
CC	1	0.743	1	0.740
RMSE	0	5.368	0	5.396
Std. Dev	8.018	5.991	8.018	6.080

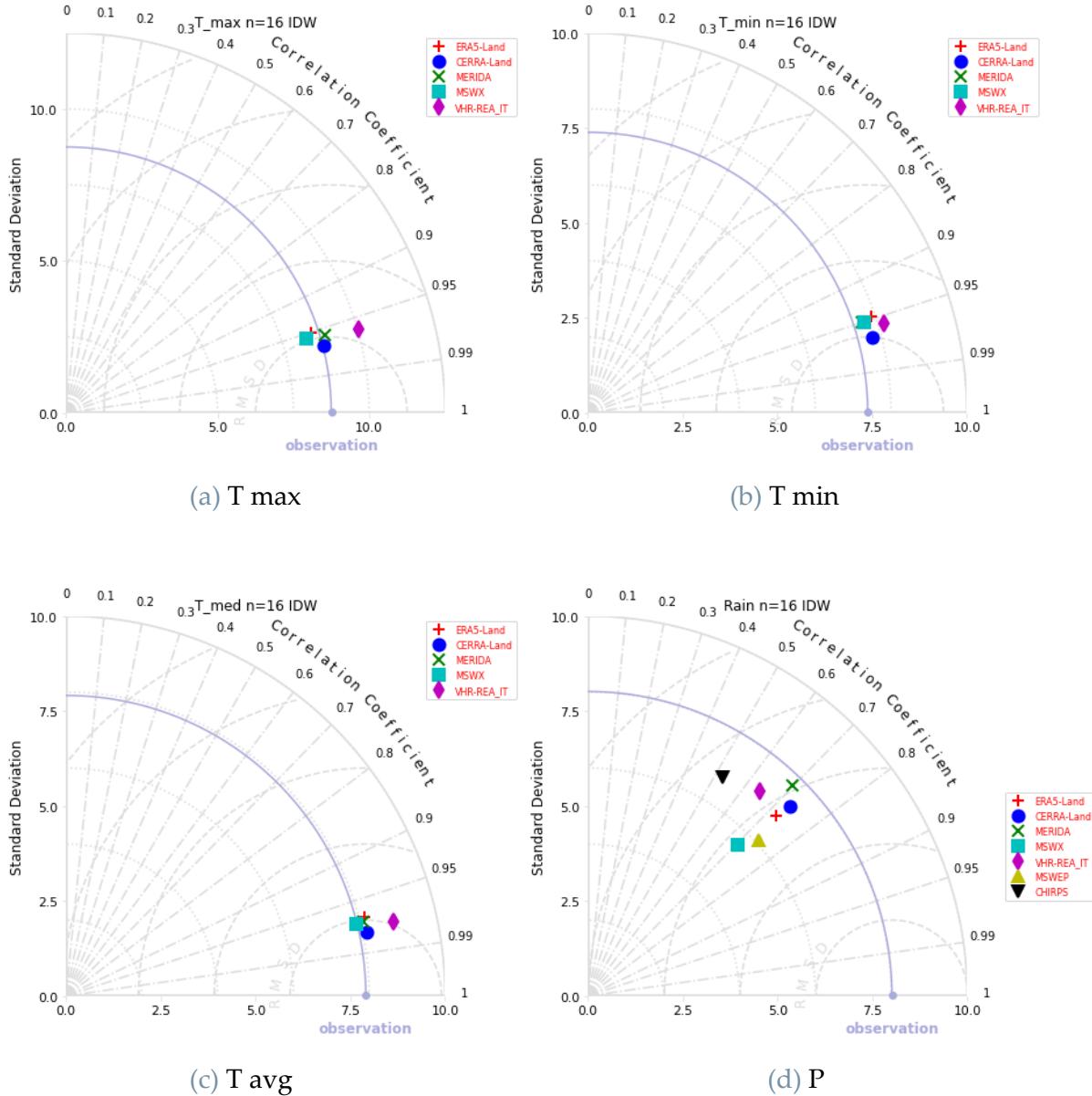


Figure 11: Taylor diagrams of the variables T max, T avg, T min and P of each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest 16 grid points, aggregated with IDW.

The addition of an altitude correction yields the most favourable outcome in terms of the similarity between the standard deviations of the two series, as shown in Figure 12 and in Table 8. However, the performance of CC and RMSE is observed to decline marginally. The most optimal reanalysis is yet still CERRA, with Era5-Land offering a second-best performance.

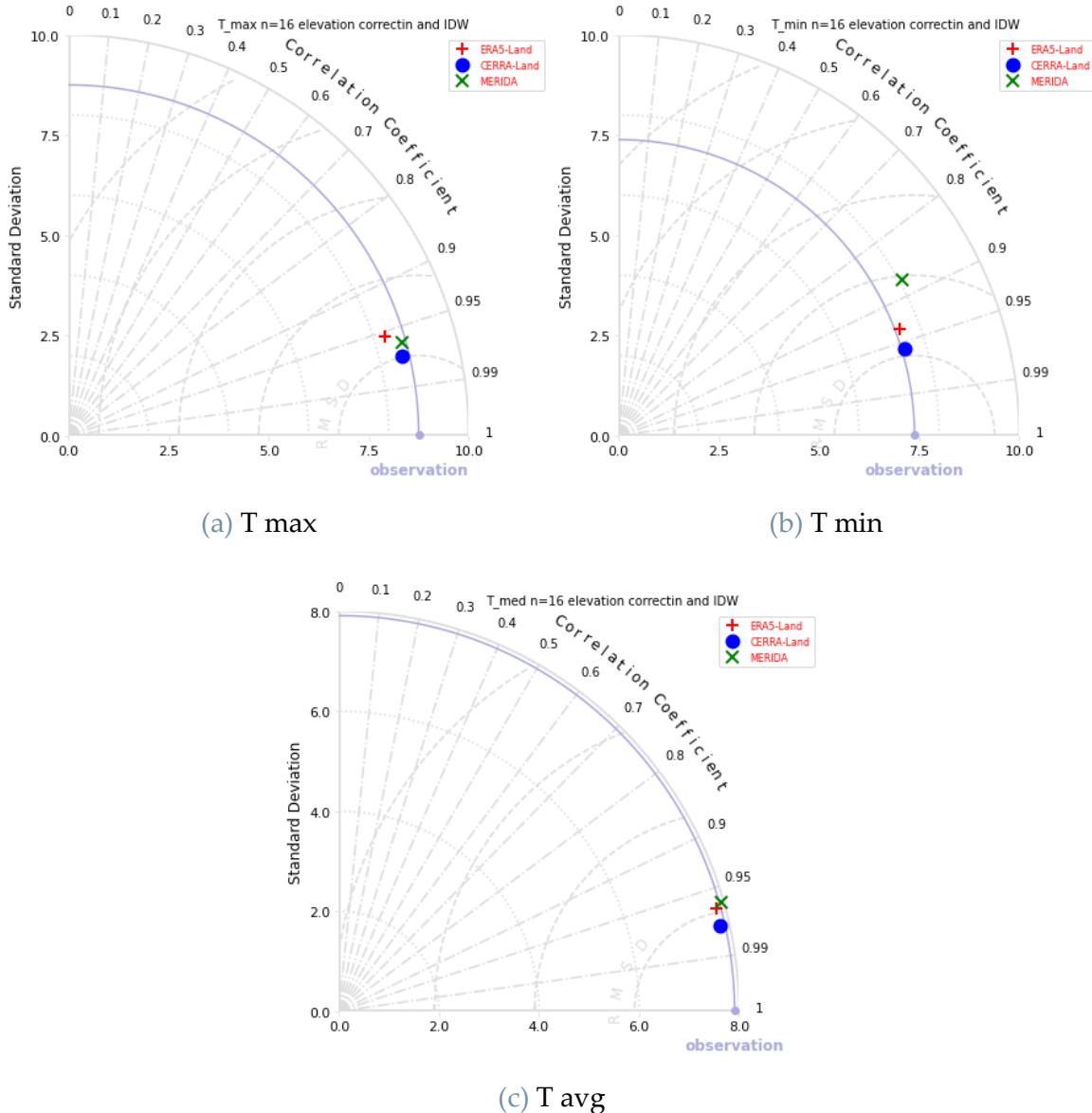


Figure 12: Taylor diagrams of the variables T max, T min and T avg of each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest 16 grid points with altitude correction and aggregated with IDW.

Table 8: Taylor diagram statistics for the CERRA reanalysis of the variable average temperature, considering sixteen points with and without elevation correction and aggregated with IDW.

T avg	N=16 IDW		N=16 elev.corr. + IDW	
	Observation	Model	Observation	Model
CC	1	0.979	1	0.975
RMSE	0	1.675	0	1.744
Std. Dev	7.869	8.118	7.869	7.804

In conclusion, with regard to temperature, the reanalysis dataset that is most consistent with the observations, closest to the reference point, is CERRA. This model demonstrates the most favourable results in terms of correlation coefficient and RMSE across all aforementioned cases. It was therefore decided to proceed with the analysis using this model.

Furthermore, Era5-Land was selected as supplementary analysis, given its extensive temperature time series data, which extends to mid-2023, a period two years longer than the time series available for CERRA. The objective was to ascertain whether this could be a factor influencing the performance of machine learning models.

With regard to precipitation, the decision was taken to proceed by testing the following models: Era5-Land, CERRA and MSWEP, that on the grounds they exhibit the most favourable performance, having the closest distance from the reference point of the observations in the Taylor diagrams. The MSWEP model demonstrated the most optimal performance, exhibiting the highest correlation coefficient and the lowest root mean square error (RMSE). The CERRA and ERA5-Land models, which exhibited a relatively close standard deviation to the observed data, were selected as the second and third best models, respectively.

The results demonstrate that the best performance is not necessarily obtained from reanalyses with higher resolutions or those downscaled specifically for the Italian area. Instead, the global reanalyses with lower resolutions in this case perform better.

## 3.2. Machine Learning Model Performance

For each variable and station, all possible combinations of the different types of models (SVR and RF) and the selected reanalysis datasets (CERRA, ERA5-Land, MSWEP) were tested in order to identify the optimal performance, with varying values of the hyper-parameters of the models.

In order to identify the most suitable model, the performance of the two indicators,  $R^2$  and RMSE, in the validation phase was considered to be of primary importance. Indeed, the models with the highest  $R^2$  and lowest RMSE in the validation phase were selected. Additionally, the time required to identify the most suitable combination of hyper-parameters was taken into account, which was influenced by two key factors: the width of the parameter search range defined in the grid search and the type of model employed. The fastest model type for optimising parameters is Support Vector Regression with a radial (RBF) kernel, followed by Random Forest and finally SVR with linear kernel.

The optimisation of the hyper-parameters through Grid Search was of paramount importance in order to obtain satisfactory performance in the reconstruction models for the maximum, mean and minimum temperature and precipitation series.

With regard to temperature, the models demonstrated consistently high accuracy, with  $R^2$  values exceeding 0.85 and RMSE values consistently below 2°C for all the stations.

The models demonstrated a superior capacity to represent mean temperature, attaining the highest  $R^2$  values (between 0.96 and 0.99) and the lowest root mean square errors (RMSEs), which were equal to or less than 1.5°C. In contrast, the reconstruction of minimum temperature exhibited the poorest performance, with  $R^2$  values between 0.85 and 0.98 and RMSE values between 1 and 2.5°C. Conversely, maximum temperature demonstrated a more favourable outcome, with  $R^2$  values between 0.90 and 0.98 and RMSE values between 1 and 2°C.

The two models, Support Vector Regression (SVR) and Random Forest (RF), demonstrated comparable and highly similar performance when the optimal parameters were identified for each station.

Similarly, the two reanalyses (CERRA and ERA5-Land) exhibited comparable performance, with minimal discrepancy. However, there was a notable divergence

observed in the case of station *trn017*, situated in Trentino Alto Adige at an elevation exceeding 2,000 meters. In this instance, the reconstruction of the minimum temperature using the ERA5-Land reanalysis yielded an  $R^2$  of 0.86 and an RMSE of  $2.5^\circ\text{C}$ , while with CERRA, the  $R^2$  was 0.94 and the RMSE was  $1.5^\circ\text{C}$ . In contrast, there was no such marked difference in the performance of the models of mean and maximum temperature.

The models for the reconstruction of the precipitation series were found to be considerably more complex than those for the temperature series. In this scenario, a greater degree of variability in performance was observed among the various stations. The incorporation of weights into the models was found to be pivotal in enhancing their predictive capacity. In the absence of these weights, the models were unable to identify an optimal function, resulting in the generation of negative  $R^2$  values in certain instances. It is notable that not all stations demonstrate consistent performance. The values of the coefficient of determination exhibit considerable variability, ranging from 0.1 to 0.8, depending on the station in question. Whereas, the RMSEs demonstrate a relatively consistent error, varying between approximately 3 and 6 millimetres. In contrast to temperature, which exhibited minimal variability, the efficacy of precipitation models is contingent upon both the specific model employed and the reanalysis data utilized as input. The only model that demonstrated an acceptable level of accuracy for all observation stations was the SVR with a radial (RBF) kernel. In contrast, utilising SVR with a linear kernel necessitated an excessive amount of computational time for the Grid Search method to identify optimal hyper-parameters, rendering the process impractical on a standard laptop within a limited timeframe. Random Forest, while demonstrating favourable outcomes at certain stations, exhibited inferior performance compared to SVR and were effective only at a very restricted number of stations.

An example of the outcomes obtained for all temperature and precipitation variables for station *lmb021* is presented in Tables 9-15.

Table 9: Comparison of performance of different type of models for mean temperature of station *lmb021*, using as input CERRA.

T avg	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
<b>SVR linear</b>	0.9946	0.5862	<b>0.9943</b>	<b>0.5973</b>
<b>SVR RBF</b>	0.9957	0.5193	0.9928	0.6821
<b>RF</b>	0.9988	0.2734	0.9938	0.6253

Table 10: Comparison of performance of different type of models of variable T max of station *lmb021*, using as input CERRA.

T max	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
<b>SVR linear</b>	0.9862	1.0237	0.9849	1.0693
<b>SVR RBF</b>	0.9856	1.0454	0.9850	1.0647
<b>RF</b>	0.9939	0.6809	<b>0.9851</b>	<b>1.0634</b>

Table 11: Comparison of performance of different type of models of variable T min of station *lmb021*, using as input CERRA.

T min	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
<b>SVR linear</b>	0.9829	0.9599	<b>0.9824</b>	<b>0.9706</b>
<b>SVR RBF</b>	0.9848	0.9046	0.9802	1.0362
<b>RF</b>	0.9936	0.5897	0.9808	1.0168

Table 12: Comparison of performance of different reanalysis of variable T avg of station *lmb021*, using the model SVR with linear kernel.

T avg	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
<b>Era5-Land</b>	0.9867	0.9213	0.9859	0.9449
<b>CERRA</b>	0.9946	0.5862	<b>0.9943</b>	<b>0.5973</b>

Table 13: Comparison of performance of different reanalysis of variable T max of station lmb021, using the model SVR with linear kernel.

T max	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
Era5-Land	0.9788	1.2715	0.9779	1.2959
CERRA	0.9862	1.0237	<b>0.9849</b>	<b>1.0693</b>

Table 14: Comparison of performance of different reanalysis of variable T min of station lmb021, using the model SVR with linear kernel.

T min	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
Era5-Land	0.9690	1.2983	0.9679	1.3187
CERRA	0.9829	0.9599	<b>0.9824</b>	<b>0.9706</b>

Table 15: Comparison of performance of different reanalysis of variable P of station lmb021, using the model SVR with radial kernel.

P	Training		Test	
	R <sup>2</sup>	RMSE [mm]	R <sup>2</sup>	RMSE [mm]
Era5-Land	0.6341	5.5054	0.6079	5.6449
CERRA	0.6262	5.7084	0.6059	5.8350
MSWEP	0.7079	4.9300	<b>0.6837</b>	<b>5.1129</b>

The SVR model with linear kernel demonstrated optimal performance in 67% of cases for maximum temperature, 81% for mean temperature and 84% for minimum temperature. With regard to reanalysis, CERRA demonstrated superior performance in 84% of cases for maximum temperature, 88% for mean temperature and 97% for minimum temperature, in comparison to ERA5-Land. The most effective reanalysis for precipitation was MSWEP, which demonstrated optimal performance in 77% of cases. The discrepancies among the reanalysis datasets used for precipitation are more pronounced than those observed for temperature. The Tables 16 and 17 illustrate the

most suitable reanalysis and model, respectively, for each station and variable, based on their performance.

Table 16: Best reanalysis dataset for each station and variable.

Code	Elevation [m]	T max	T avg	T min	P
abr034	685	CERRA	CERRA	CERRA	ERA5-Land
abr047	85	CERRA	CERRA	CERRA	CERRA
bsl008	455	ERA5-Land	CERRA	CERRA	ERA5-Land
clb002	173	CERRA	ERA5-Land	CERRA	MSWEP
cmp015	69	ERA5-Land	CERRA	CERRA	MSWEP
ero001	1118	CERRA	CERRA	CERRA	CERRA
ero002	31	CERRA	CERRA	CERRA	MSWEP
ero009	38	CERRA	CERRA	CERRA	MSWEP
fvg009	35	CERRA	CERRA	CERRA	MSWEP
laz013	13	CERRA	CERRA	CERRA	MSWEP
laz026	320	ERA5-Land	CERRA	CERRA	MSWEP
laz033	362	CERRA	CERRA	CERRA	MSWEP
lig009	130	CERRA	CERRA	CERRA	CERRA
lmb002	750	CERRA	CERRA	CERRA	MSWEP
lmb015	28	CERRA	CERRA	CERRA	CERRA
lmb021	224	CERRA	CERRA	CERRA	MSWEP
lmb039	68	CERRA	CERRA	CERRA	MSWEP
lmb084	427	CERRA	CERRA	CERRA	MSWEP
lmb255	1386	CERRA	CERRA	CERRA	MSWEP
mbr001	455	ERA5-Land	CERRA	CERRA	MSWEP
mbr006	300	CERRA	ERA5-Land	CERRA	MSWEP
mbr026	350	ERA5-Land	ERA5-Land	CERRA	MSWEP
mrc009	25	CERRA	CERRA	CERRA	MSWEP
pgl005	195	CERRA	CERRA	CERRA	MSWEP
pgl008	76	CERRA	ERA5-Land	CERRA	MSWEP
pgl012	106	ERA5-Land	CERRA	CERRA	MSWEP
pgl040	20	CERRA	CERRA	CERRA	ERA5-Land
pmn016	229	CERRA	CERRA	CERRA	MSWEP
pmn033	170	CERRA	CERRA	CERRA	CERRA
pmn036	686	CERRA	CERRA	CERRA	ERA5-Land
pmn074	200	CERRA	CERRA	CERRA	ERA5-Land
scl007	430	CERRA	CERRA	ERA5-Land	MSWEP
scl014	170	CERRA	CERRA	CERRA	MSWEP

Code	Elevation [m]	T max	T avg	T min	P
scl073	95	CERRA	CERRA	CERRA	MSWEP
sdr013	339	CERRA	CERRA	CERRA	MSWEP
trn009	202	CERRA	CERRA	CERRA	MSWEP
trn017	2450	CERRA	CERRA	CERRA	MSWEP
tsc001	2	CERRA	CERRA	CERRA	MSWEP
tsc002	595	ERA5-Land	ERA5-Land	CERRA	MSWEP
tsc003	60	CERRA	CERRA	CERRA	MSWEP
vda006	1500	CERRA	CERRA	CERRA	MSWEP
vnt023	90	CERRA	CERRA	CERRA	MSWEP
vnt160	2	CERRA	CERRA	CERRA	MSWEP

Table 17: Best ML model for each station and variable.

Code	Elevation [m]	T max	T avg	T min	P
abr034	685	SVR .RBF	SVR .RBF	SVR linear	SVR .RBF
abr047	85	SVR linear	SVR linear	SVR linear	SVR .RBF
bsl008	455	SVR linear	SVR linear	SVR linear	SVR .RBF
clb002	173	SVR linear	SVR linear	SVR linear	SVR .RBF
cmp015	69	SVR linear	SVR linear	RF	SVR .RBF
ero001	1118	SVR linear	SVR linear	RF	SVR .RBF
ero002	31	SVR linear	SVR linear	SVR linear	SVR .RBF
ero009	38	RF	SVR linear	SVR linear	SVR .RBF
fvg009	35	SVR linear	SVR linear	SVR linear	SVR .RBF
laz013	13	RF	SVR linear	SVR linear	SVR .RBF
laz026	320	SVR linear	SVR linear	SVR linear	SVR .RBF
laz033	362	SVR linear	RF	SVR linear	SVR .RBF
lig009	130	RF	SVR linear	SVR linear	SVR .RBF
lmb002	750	RF	SVR linear	SVR linear	SVR .RBF
lmb015	28	SVR linear	SVR linear	SVR linear	SVR .RBF
lmb021	224	RF	SVR linear	SVR linear	SVR .RBF
lmb039	68	SVR linear	SVR linear	SVR linear	SVR .RBF
lmb084	427	SVR linear	SVR linear	SVR linear	SVR .RBF
lmb255	1386	RF	SVR linear	SVR linear	SVR .RBF
mbr001	455	SVR linear	SVR linear	SVR linear	SVR .RBF
mbr006	300	SVR linear	SVR linear	SVR linear	SVR .RBF
mbr026	350	SVR linear	SVR linear	SVR linear	SVR .RBF

Code	Elevation [m]	T max	T avg	T min	P
mrc009	25	SVR linear	SVR linear	SVR linear	SVR .RBF
pgl005	195	SVR linear	SVR linear	SVR linear	SVR .RBF
pgl008	76	SVR linear	SVR linear	SVR linear	SVR .RBF
pgl012	106	SVR linear	SVR linear	SVR linear	SVR .RBF
pgl040	20	SVR linear	SVR linear	SVR linear	SVR .RBF
pmn016	229	SVR linear	SVR linear	SVR linear	SVR .RBF
pmn033	170	RF	SVR linear	RF	SVR .RBF
pmn036	686	RF	RF	RF	SVR .RBF
pmn074	200	SVR linear	SVR linear	SVR linear	SVR .RBF
scl007	430	SVR linear	SVR .RBF	SVR linear	SVR .RBF
scl014	170	RF	SVR linear	SVR linear	SVR .RBF
scl073	95	SVR linear	SVR linear	SVR linear	SVR .RBF
sdr013	339	SVR linear	SVR linear	SVR linear	SVR .RBF
trn009	202	RF	SVR .RBF	RF	SVR .RBF
trn017	2450	RF	SVR .RBF	SVR .RBF	SVR .RBF
tsc001	2	SVR linear	SVR linear	SVR linear	SVR .RBF
tsc002	595	RF	SVR .RBF	SVR linear	SVR .RBF
tsc003	60	RF	SVR linear	SVR linear	SVR .RBF
vda006	1500	SVR linear	SVR linear	RF	SVR .RBF
vnt023	90	SVR linear	RF	SVR linear	SVR .RBF
vnt160	2	SVR linear	SVR linear	SVR linear	SVR .RBF

The results were not found to be influenced by any particular spatial distribution of the stations. The stations for which a reanalysis or a model performs best are scattered without an obvious pattern, as can be seen from the Figures 13-15.

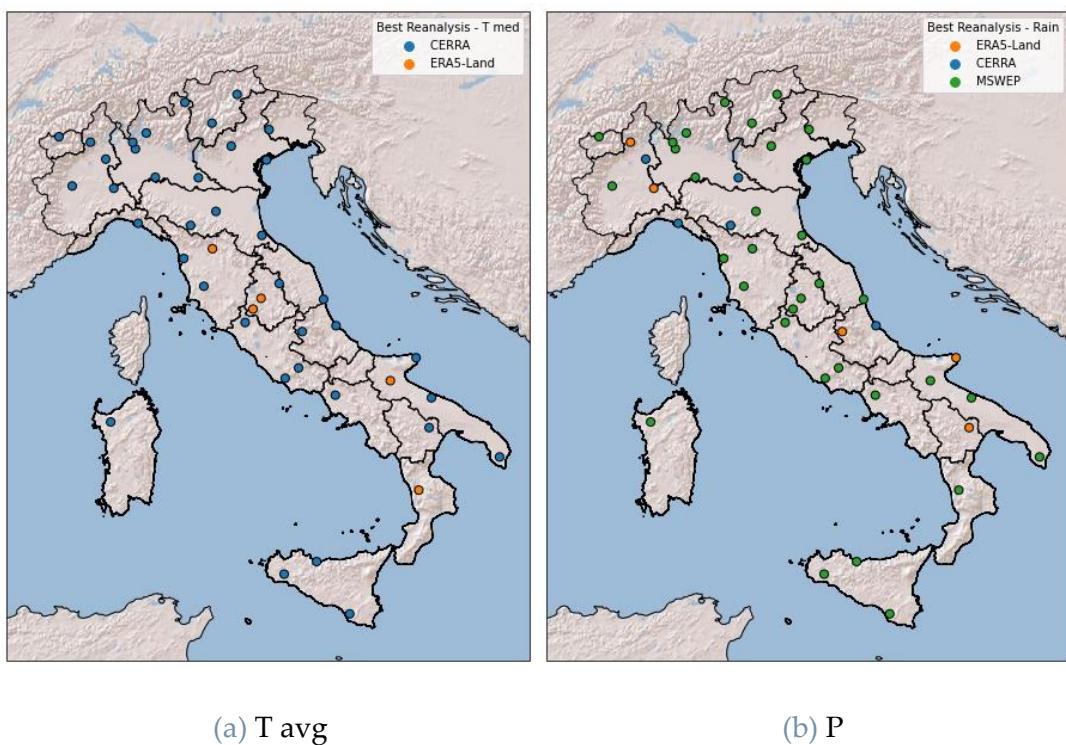


Figure 13: Spatial distribution of the best resulting reanalysis dataset for precipitation and mean temperature.

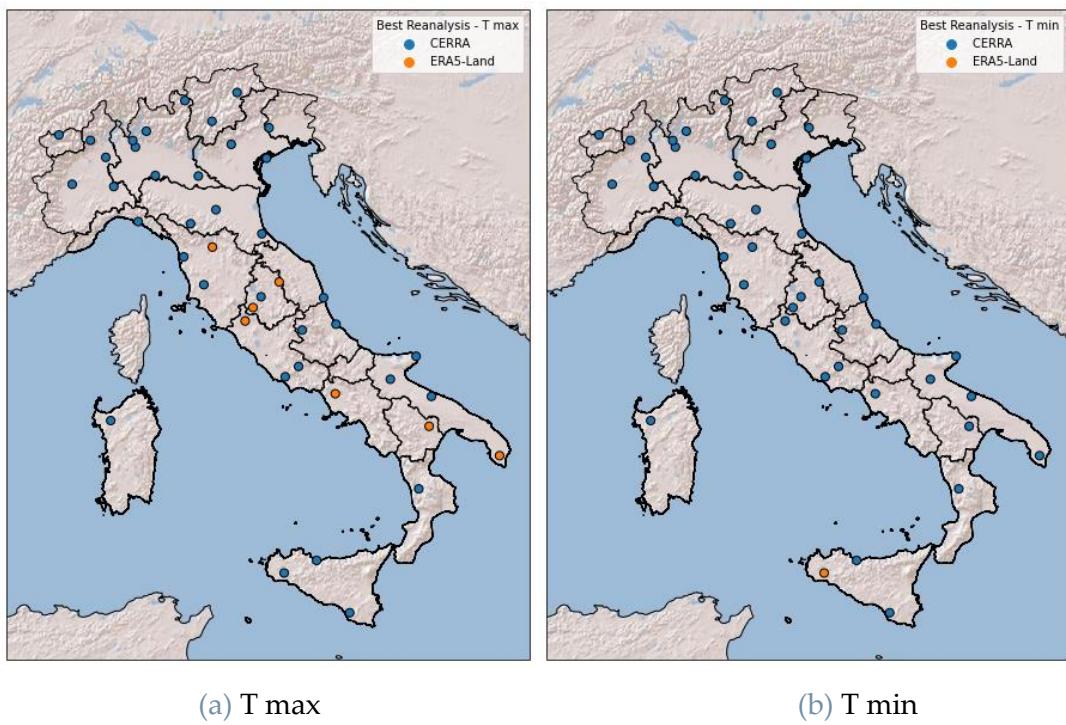


Figure 14: Spatial distribution of the best resulting reanalysis dataset for maximum and minimum temperature.

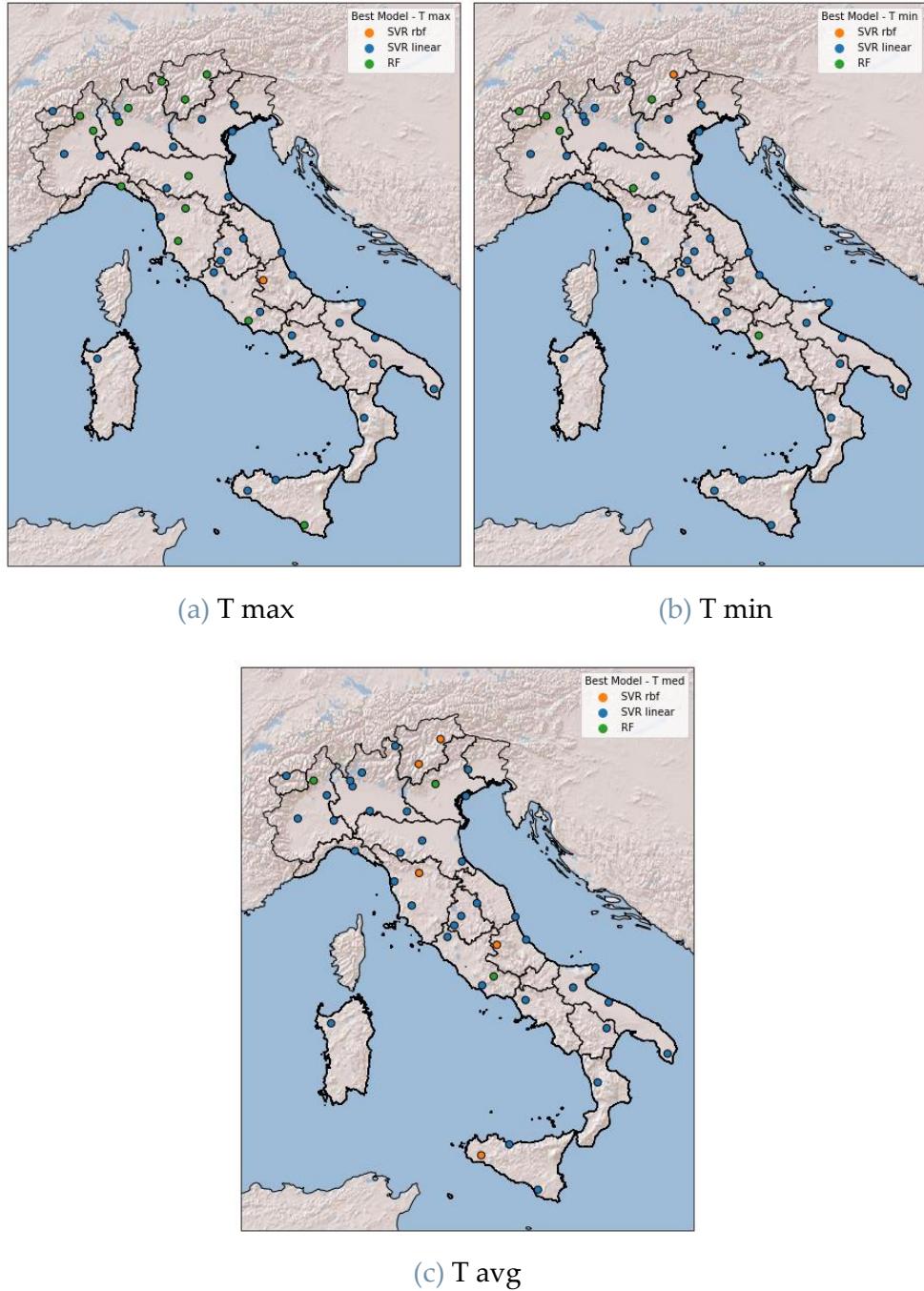


Figure 15: Spatial distribution of the best resulting ML model for maximum, minimum and mean temperature.

In order reconstruct the temperature and precipitation series to calculate the climatologies, it was opted to select input reanalyses and models that demonstrated superior performance on average across the entire sample of stations. Accordingly, the temperature series were reconstructed using SVR with a linear kernel and the CERRA reanalysis, whereas the precipitation series were reconstructed using SVR with a radial

kernel and the MSWEP reanalysis, based on the average performance obtained at all stations. The Tables 18-21 demonstrate the degree of accuracy achieved by each model employed for the reconstruction of the series for each variable and each station.

Table 18: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable average temperature for each station.

T avg	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
abr034	0.9763	1.1153	0.9704	1.2150
abr047	0.9878	0.7119	0.9849	0.7675
bsl008	0.9864	0.8587	0.9852	0.8910
clb002	0.9745	1.1291	0.9722	1.1736
cmp015	0.9872	0.7233	0.9855	0.7677
ero001	0.9845	0.8820	0.9813	0.9525
ero002	0.9946	0.5707	0.9942	0.5859
ero009	0.9912	0.7468	0.9893	0.8011
fvg009	0.9929	0.6469	0.9924	0.6665
laz013	0.9904	0.6189	0.9893	0.6479
laz026	0.9945	0.4858	0.9932	0.5387
laz033	0.9895	0.7136	0.9874	0.7752
lig009	0.9893	0.5999	0.9872	0.6515
lmb002	0.9910	0.6843	0.9902	0.7063
lmb015	0.9879	0.9401	0.9845	1.0581
lmb021	0.9946	0.5862	0.9943	0.5973
lmb039	0.9942	0.6548	0.9936	0.6818
lmb084	0.9900	0.7324	0.9889	0.7659
lmb255	0.9847	0.8596	0.9819	0.9186
mbr001	0.9861	0.8447	0.9836	0.9108
mbr006	0.9839	0.9394	0.9812	1.0074
mbr026	0.9879	0.7730	0.9812	0.9558
mrc009	0.9917	0.5991	0.9906	0.6279
pgl005	0.9928	0.5963	0.9922	0.6262
pgl008	0.9830	0.9451	0.9793	1.0389
pgl012	0.9883	0.7132	0.9865	0.7496
pgl040	0.9902	0.6186	0.9878	0.6861
pmn016	0.9889	0.8863	0.9882	0.9090
pmn033	0.9913	0.7488	0.9901	0.7895
pmn036	0.9817	0.9952	0.9791	1.0532

T avg	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
pmn074	0.9918	0.7186	0.9906	0.7539
scl007	0.9887	0.6995	0.9877	0.7312
scl014	0.9880	0.6728	0.9863	0.7191
scl073	0.9877	0.6388	0.9856	0.6795
sdr013	0.9837	0.8054	0.9819	0.8415
trn009	0.9695	1.4428	0.9654	1.5234
trn017	0.9700	1.1888	0.9657	1.2610
tsc001	0.9918	0.6054	0.9905	0.6509
tsc002	0.9829	0.8974	0.9765	1.0342
tsc003	0.9923	0.6393	0.9917	0.6687
vda006	0.9760	1.0088	0.9721	1.0669
vnt023	0.9897	0.8281	0.9878	0.9004
vnt160	0.9953	0.5370	0.9941	0.6007

Table 19: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable minimum temperature for each station.

T min	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
abr034	0.9400	1.5783	0.9282	1.6763
abr047	0.9677	1.1268	0.9634	1.1626
bsl008	0.9505	1.4682	0.9472	1.5083
clb002	0.9352	1.6028	0.9316	1.6411
cmp015	0.9552	1.2925	0.9401	1.4865
ero001	0.9750	1.027	0.9723	1.0669
ero002	0.9800	0.9791	0.9793	0.9926
ero009	0.9706	1.2046	0.9677	1.254
fvg009	0.9693	1.3001	0.9672	1.3329
laz013	0.9667	1.0869	0.9638	1.1239
laz026	0.9785	0.8841	0.9769	0.9083
laz033	0.9812	0.8282	0.9799	0.8484
lig009	0.9761	0.8668	0.9742	0.892
lmb002	0.9770	1.0031	0.9758	1.0241
lmb015	0.9780	1.1381	0.9749	1.212
lmb021	0.9829	0.9599	0.9824	0.9706
lmb039	0.9815	1.047	0.9797	1.0884
lmb084	0.9753	1.0273	0.9737	1.0551

T min	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
lmb255	0.9700	1.0761	0.9666	1.1098
mbr001	0.9462	1.4521	0.9419	1.5004
mbr006	0.9644	1.2059	0.962	1.2493
mbr026	0.9657	1.1154	0.9619	1.1613
mrc009	0.9770	0.9664	0.9755	0.9664
pgl005	0.9811	0.8991	0.98	0.918
pgl008	0.9495	1.4883	0.9432	1.5765
pgl012	0.9572	1.291	0.9543	1.3154
pgl040	0.9776	0.923	0.976	0.9525
pmn016	0.9774	1.175	0.9764	1.1976
pmn033	0.9733	1.2696	0.9698	1.3325
pmn036	0.9699	1.1456	0.9681	1.1721
pmn074	0.9791	1.0356	0.9772	1.06
scl007	0.9656	1.0847	0.964	1.1069
scl014	0.9719	0.982	0.9688	1.0326
scl073	0.9728	0.9019	0.9671	0.9746
sdr013	0.9637	1.0416	0.9603	1.0705
trn009	0.9532	1.6505	0.9481	1.7237
trn017	0.9494	1.5339	0.9441	1.5965
tsc001	0.9737	1.0279	0.9718	1.0546
tsc002	0.9547	1.2961	0.9400	1.4570
tsc003	0.978	0.9557	0.9768	0.9747
vda006	0.9619	1.1537	0.9573	1.1969
vnt023	0.9658	1.3984	0.9625	1.4379
vnt160	0.9833	0.9608	0.9824	0.9774

Table 20: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable maximum temperature for each station.

T max	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
abr034	0.9712	1.4255	0.9656	1.5227
abr047	0.9718	1.1155	0.9693	1.1388
bsl008	0.9824	1.1296	0.9812	1.1591
clb002	0.9694	1.4401	0.9669	1.4925
cmp015	0.977	1.0735	0.9745	1.1283

T max	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
ero001	0.976	1.2172	0.9733	1.2755
ero002	0.9826	1.1257	0.9816	1.1478
ero009	0.9831	1.2277	0.9783	1.2866
fvg009	0.9844	1.0144	0.9833	1.0465
laz013	0.985	0.8253	0.9840	0.8463
laz026	0.9841	0.9489	0.9818	1.003
laz033	0.9785	1.2112	0.9761	1.2656
lig009	0.9762	0.9361	0.9742	0.9647
lmb002	0.9835	1.0330	0.9824	1.0596
lmb015	0.9794	1.4345	0.9767	1.5251
lmb021	0.9862	1.0237	0.9849	1.0693
lmb039	0.9841	1.2407	0.9820	1.3074
lmb084	0.9845	1.0242	0.9828	1.0684
lmb255	0.9506	1.8070	0.9385	1.9718
mbr001	0.9783	1.2591	0.9740	1.3630
mbr006	0.9705	1.4944	0.9686	1.5408
mbr026	0.9769	1.3057	0.9703	1.4585
mrc009	0.9811	0.9251	0.9801	0.9415
pgl005	0.9838	0.9531	0.9826	0.9790
pgl008	0.9793	1.1933	0.9777	1.2329
pgl012	0.9847	0.9129	0.9812	0.9865
pgl040	0.9781	0.9552	0.9755	1.0024
pmn016	0.9781	1.3476	0.9755	1.4181
pmn033	0.9792	1.2422	0.9775	1.2860
pmn036	0.9624	1.5765	0.9591	1.6266
pmn074	0.9819	1.2365	0.9792	1.3133
scl007	0.9828	0.9920	0.9818	1.0152
scl014	0.9781	0.9773	0.9759	1.0227
scl073	0.9747	0.9946	0.9705	1.0578
sdr013	0.9796	1.0262	0.9779	1.0553
trn009	0.9528	2.0229	0.9482	2.0984
trn017	0.9581	1.4714	0.9545	1.5154
tsc001	0.9816	0.9917	0.9802	1.0198
tsc002	0.9782	1.1810	0.9743	1.2672
tsc003	0.9802	1.1950	0.9789	1.2222
vda006	0.9585	1.5381	0.9536	1.5943
vnt023	0.9838	1.1202	0.9820	1.1750

T max	Training		Test	
	R <sup>2</sup>	RMSE [°C]	R <sup>2</sup>	RMSE [°C]
vnt160	0.9828	1.0664	0.9799	1.1442

Table 21: Performance of the SVR models with radial kernel and MSWEP as input, used to reconstruct the synthetic series of the variable precipitation for each station.

P	Training		Test	
	R <sup>2</sup>	RMSE [mm]	R <sup>2</sup>	RMSE [mm]
abr034	0.3386	4.5126	0.3888	4.6075
abr047	0.6028	5.7939	0.3232	6.6960
bsl008	0.5182	4.5973	0.5031	4.6748
clb002	0.6979	3.7311	0.6865	3.7890
cmp015	0.6631	4.5284	0.6483	4.6080
ero001	0.2758	5.9423	0.1342	6.3118
ero002	0.6882	3.3576	0.6875	3.4438
ero009	0.6534	4.3653	0.6255	4.4833
fvg009	0.72637	6.25615	0.7251	6.3545
laz013	0.6329	3.5984	0.6129	3.6906
laz026	0.5707	5.2960	0.5746	5.3159
laz033	0.7359	5.2429	0.7273	5.3035
lig009	0.5725	6.4197	0.5373	6.5898
lmb002	0.7483	6.5756	0.7300	6.7674
lmb015	0.4940	4.3776	0.4783	4.3936
lmb021	0.7203	4.8246	0.6743	5.1732
lmb039	0.6384	3.5235	0.6214	3.5828
lmb084	0.6906	6.1355	0.6703	6.2844
lmb255	0.7349	3.0595	0.7104	3.1805
mbr001	0.5675	6.4758	0.5585	6.5771
mbr006	0.4972	4.3679	0.4906	4.3977
mbr026	0.5916	3.6005	0.5832	3.6968
mrc009	0.4325	5.1759	0.4352	5.2438
pgl005	0.5109	3.7143	0.5001	3.8018
pgl008	0.4938	3.7872	0.4958	3.8558
pgl012	0.5789	4.2378	0.5707	4.2597
pgl040	0.4373	3.6874	0.4264	3.7774
pmn016	0.6095	5.1199	0.6199	5.1671
pmn033	0.6380	4.5276	0.5979	4.7152

P	Training		Test	
	R <sup>2</sup>	RMSE [mm]	R <sup>2</sup>	RMSE [mm]
pmn036	0.6777	8.1121	0.6449	8.4736
pmn074	0.5793	4.3468	0.5590	4.5146
scl007	0.6904	4.8950	0.6829	5.0176
scl014	0.7185	3.0175	0.7183	3.1151
scl073	0.4831	5.0170	0.4681	5.1079
sdr013	0.5471	3.9981	0.5497	4.0117
trn009	0.8180	3.6576	0.7894	3.8703
trn017	0.3328	2.8579	0.2871	2.9188
tsc001	0.8194	2.9792	0.8051	3.0682
tsc002	0.5769	5.8030	0.5601	5.9212
tsc003	0.6300	4.0717	0.5981	4.2700
vda006	0.4846	3.2528	0.4285	3.3648
vnt023	0.7127	4.4996	0.6747	4.7487
vnt160	0.5994	3.4804	0.5846	3.5210

The most optimal outcome among the models employed for the reconstruction of the series concerning temperatures was observed for station *lmb021*, exhibiting a notably elevated R<sup>2</sup> of 0.9943 for the mean temperature. In contrast, the poorest performance was observed for the station *abr034* for the minimum temperature variable, with an R<sup>2</sup> in validation of 0.9282.

With regard to precipitation, the most accurate station was *tsc001*, with an R<sup>2</sup> of 0.8051. Conversely, the least accurate was *ero001*, with an R<sup>2</sup> of just 0.1341. Despite numerous attempts at optimisation, it was not possible to significantly improve the results for some stations, with *ero001* in particular showing very low performance. In the case of 11 stations (26% of the total), it was not possible to obtain a model that exceeded the performance of R<sup>2</sup> equal to 0.5 in the validation process.

It is noteworthy that some of the stations with the poorest performance are located at altitudes above 1,000 metres, such as *vda009*, *ero001* and *trn017*. However, other stations at similar altitudes, such as *lmb255*, exhibited significantly better results (R<sup>2</sup> = 0.7104). Consequently, altitude is not the only factor that can lead to poor model

performance. Other factors may also contribute to the observed deficiencies in model performance. Moreover, the length of the time series does not appear to be a significant determining factor. For instance, *ero001* has almost 17 years of observations. Other stations with fewer days of observation still achieved favourable results e.g., *clb002* that with 13.5 years of observation obtains an  $R^2$  of 0.68. These factors suggest that the poor performance of some stations may be attributable to data quality issues associated with the observations themselves, or to the inability of reanalysis datasets to adequately resolve localised phenomena such as thunderstorms, which is likely to have contributed to the poorer performance of precipitation models.

### 3.3. Climatologies

The climatologies of the reconstructed synthetic series obtained are presented in the following section, examples are shown in Figure 16 and 17. To view the complete climatological series for all 43 observation stations, please refer to the Appendix.

The climatologies presented here are representative of a lowland and a mountain station, respectively. This is evident from the temperature trend, which indicates a 5°C upward shift for *lmb021* compared to the *lmb255* station, and from the rainfall trend, which shows the highest precipitation levels for the lowland one in November, in contrast to the mountain station, where the peak occurs in typically in August.

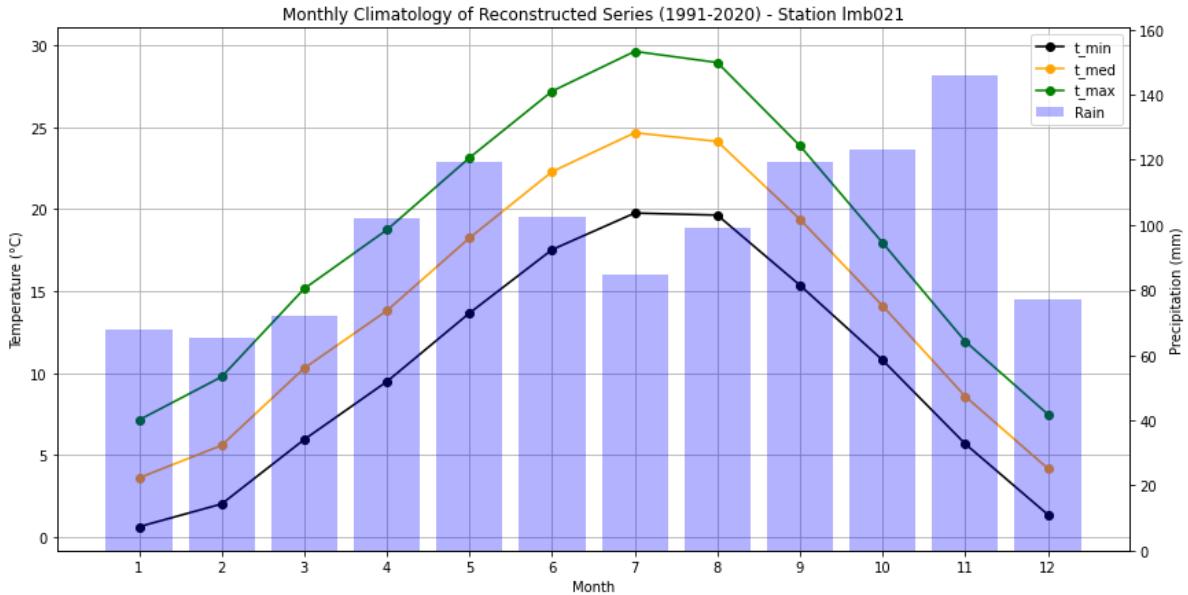


Figure 16: Monthly climatologies of the period 1991-2020 of the reconstructed synthetic series of station lmb021.

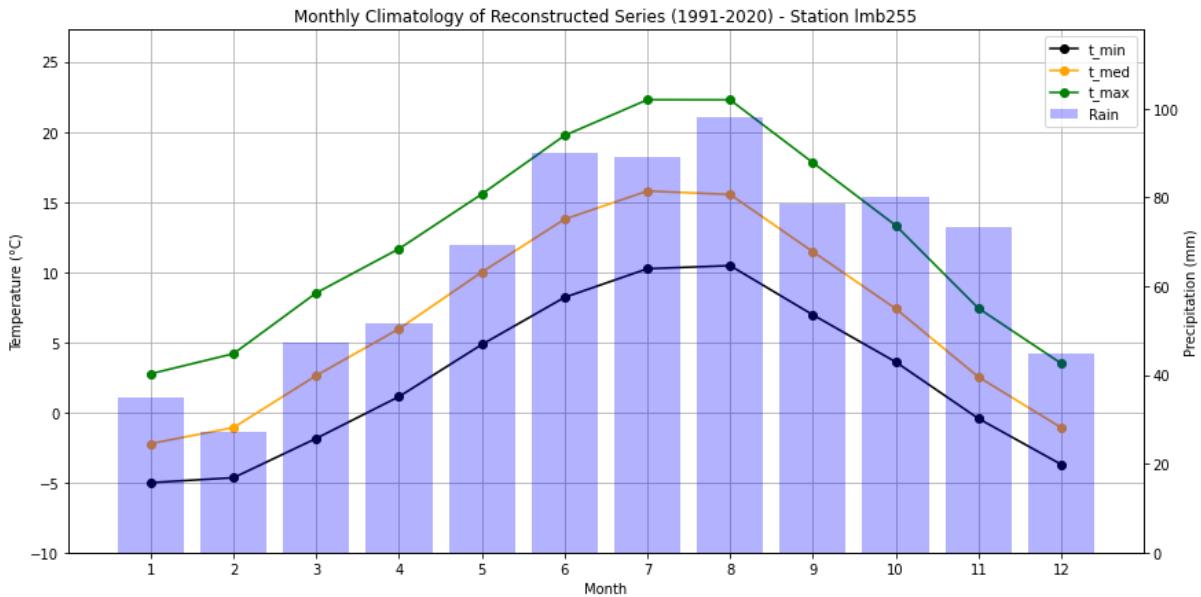


Figure 17: Monthly climatologies of the period 1991-2020 of the reconstructed synthetic series of station lmb255.

In order to ascertain the extent of the discrepancy in the climatologies resulting from the utilisation of different inputs or the employment of distinct ML modelling techniques, comparative graphical representations were devised, as illustrated in Figures 18-21. With regard to the temperature climatologies, it can be observed that the final result remains largely unaffected by the choice of the reanalysis, demonstrating a high degree of consistency. The comparable performances of the

various models yield almost identical climatologies. In the case of the lowland station (*lmb021*), the climatologies exhibit a high degree of overlap. In contrast, the mountain station exhibits a slight discrepancy, particularly in the summer months of the maximum temperature climatology and the spring months of the minimum temperature climatology. However, the divergence remains within a few degrees.

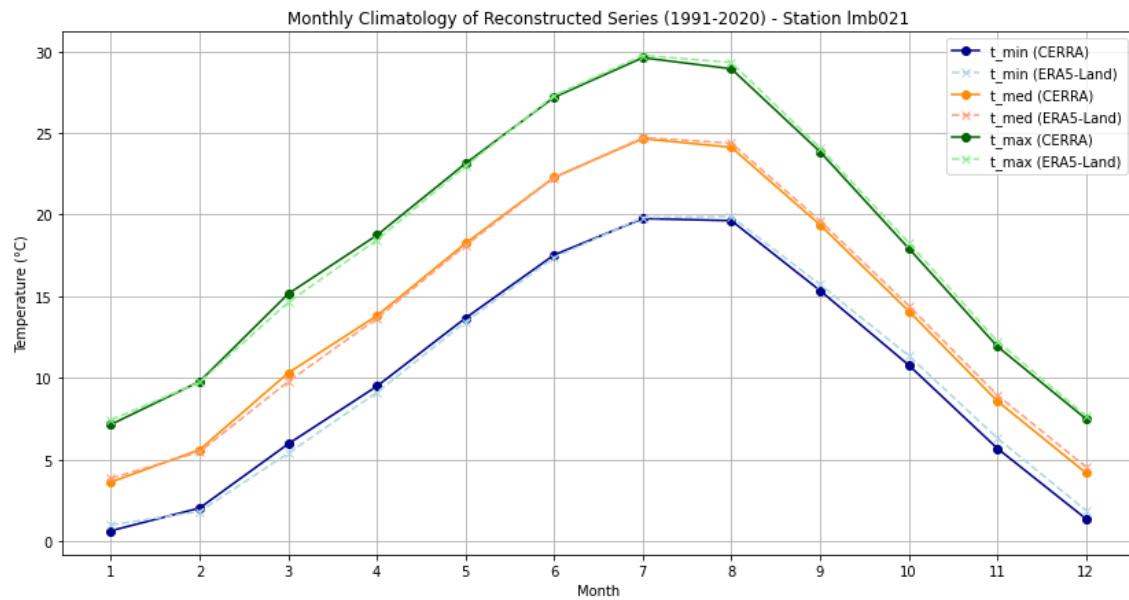


Figure 18: Comparison between monthly climatologies of temperature reconstructed with SVR with linear kernel, using as input CERRA and ERA5-Land, for the station lmb021.

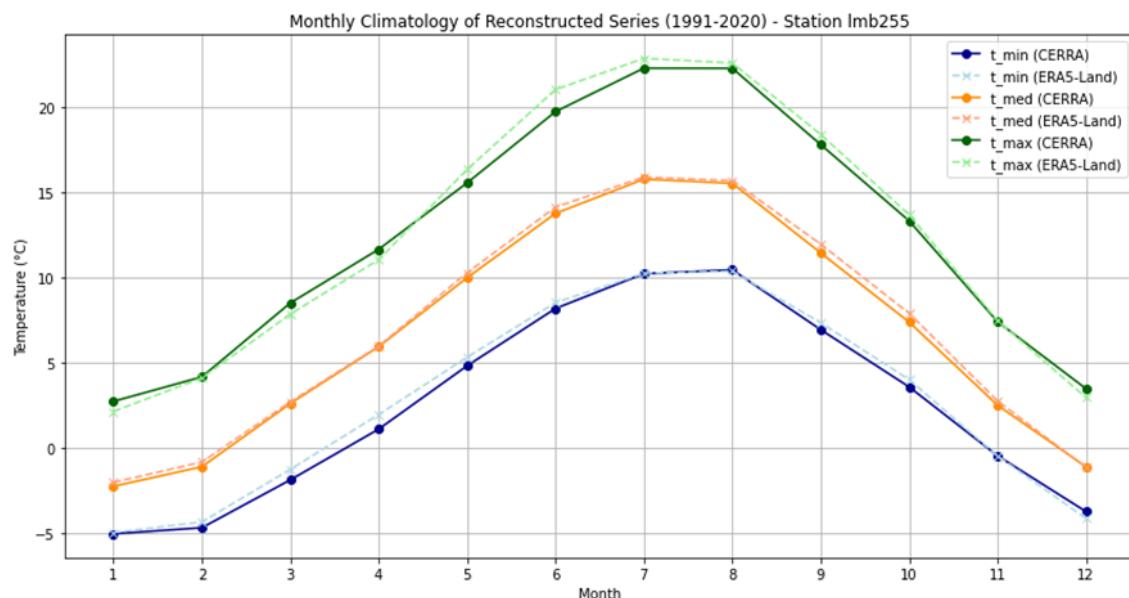


Figure 19: Comparison between monthly climatologies of temperature reconstructed with SVR with linear kernel, using as input CERRA and ERA5-Land, for the station lmb255.

Furthermore, the selection of the ML model for the lowland station exerts an even lesser influence than the choice of input, given that all climatologies are perfectly overlapping, as can be seen in Figure 20, because the performance of the SVR and RF models were still very similar. Additionally, Figure 21 illustrates that for the mountain station the climatologies of maximum temperatures derived from different models exhibit discernible discrepancies, particularly during the winter months at the end of the year and to a lesser extent during the period between April and July. In contrast, the climatologies of average temperatures and minima exhibit minimal deviation and overlap.

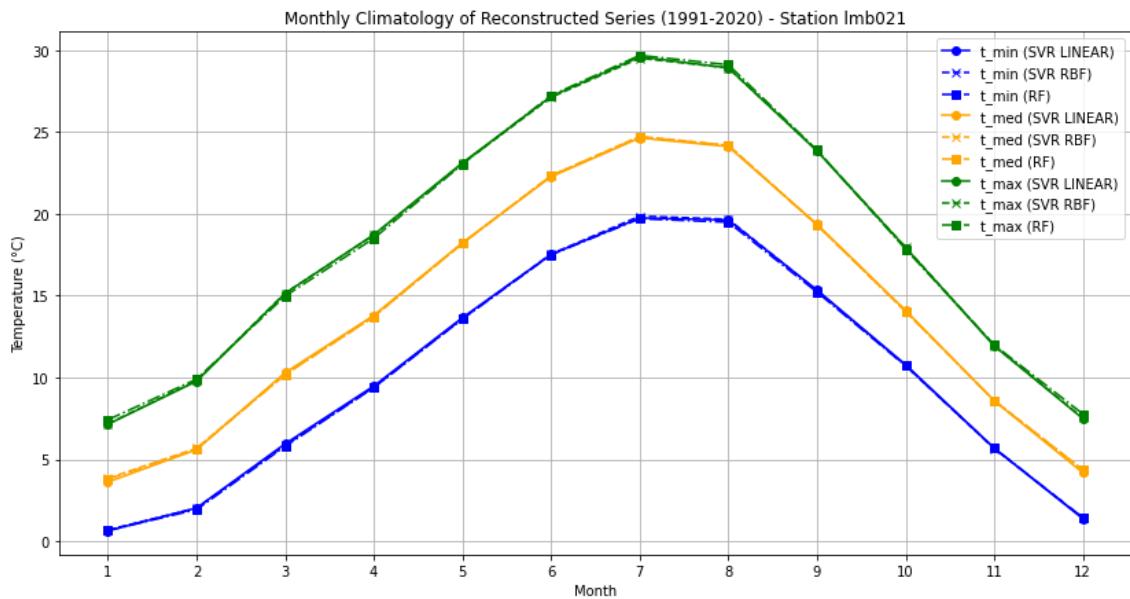


Figure 20: Comparison among monthly climatologies of temperature reconstructed with CERRA as input, using different types of machine learning models for station lmb021.

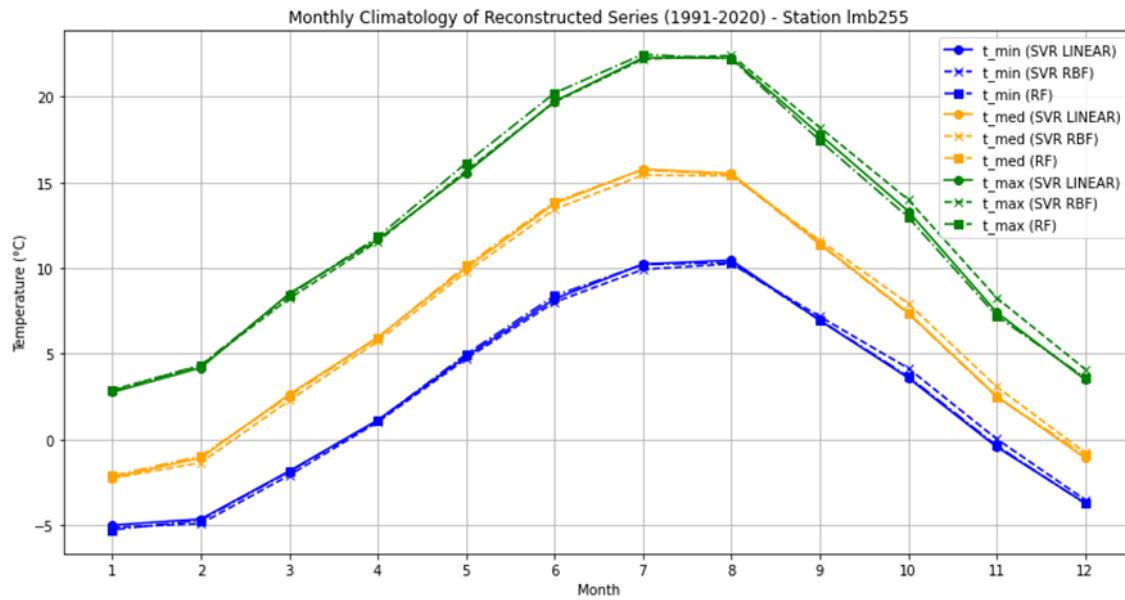


Figure 21: Comparison among monthly climatologies of temperature reconstructed with CERRA as input, using different types of machine learning models for station lmb255.

Upon examination of the precipitation climatologies displayed in Figures 22 and 23, it becomes evident that the selection of reanalysis dataset utilized as input exerts a significant influence on the monthly mean cumulative rainfall outcome, with deviations ranging up to nearly 20 millimetres in the most unfavourable months at both stations.

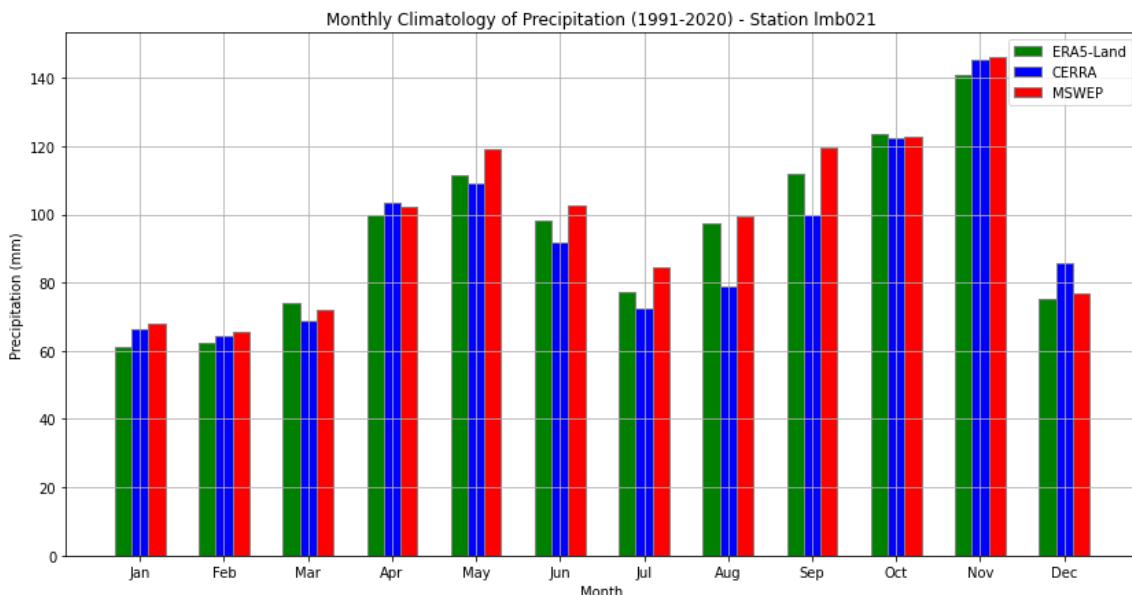


Figure 22: Comparison among monthly climatologies of precipitation reconstructed with SVR with radial kernel, using different types of reanalysis datasets as input, for lmb021.

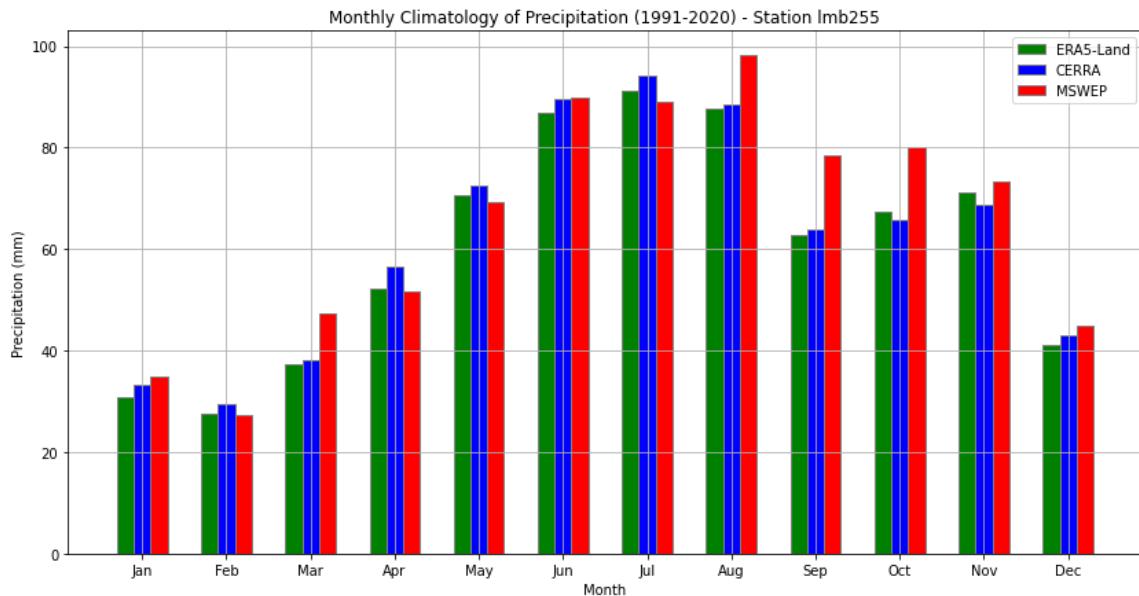


Figure 23: Comparison among monthly climatologies of precipitation reconstructed with SVR with radial kernel, using different types of reanalysis datasets as input, for lmb255.

In order to ascertain the accuracy of the reconstructed climatologies and to gain insight into whether the models selected as the most optimal and the reanalyses chosen as input allow for the reconstruction of climatologies that are in close proximity to the actual climatic conditions, a comparison was conducted between them and the climatologies calculated using only the available daily observations, despite the presence of missing data. It is evident that a comparison can only be made between the reconstructed series and the original observation series during the period of overlap. Although this comparison is not strictly accurate, as the calibration and validation of the models occurred during this common period, which was used as reference, it is the only means of determining, in addition to the model accuracy indicators, whether the reconstructed series are sufficiently similar to the observed series.

To gain insight into the range of deviation observed in the climatologies relative to the actual series, four illustrative stations were selected. Two of the selected stations demonstrated the most favourable performance with respect to temperature and

precipitation models, respectively, while the other two exhibited the least favourable performance.

With regard to temperature, the station with the most accurate model was *lmb021*. The corresponding climatologies and the difference between them and the climatologies calculated from observations are shown in Figure 24. In contrast, the station with the least accurate model was *abr034*, whose climatologies are shown in Figure 25. It can be observed that even in the most unfavourable circumstances, the monthly climatologies exhibit minimal deviation of few degrees from the observed data, thereby demonstrating their efficacy as a representation of the climatological averages. It was anticipated that the model would demonstrate a high level of resemblance during the calibration and validation period. It may therefore be reasonably assumed that the results will be favourable also in periods where there is no overlap.

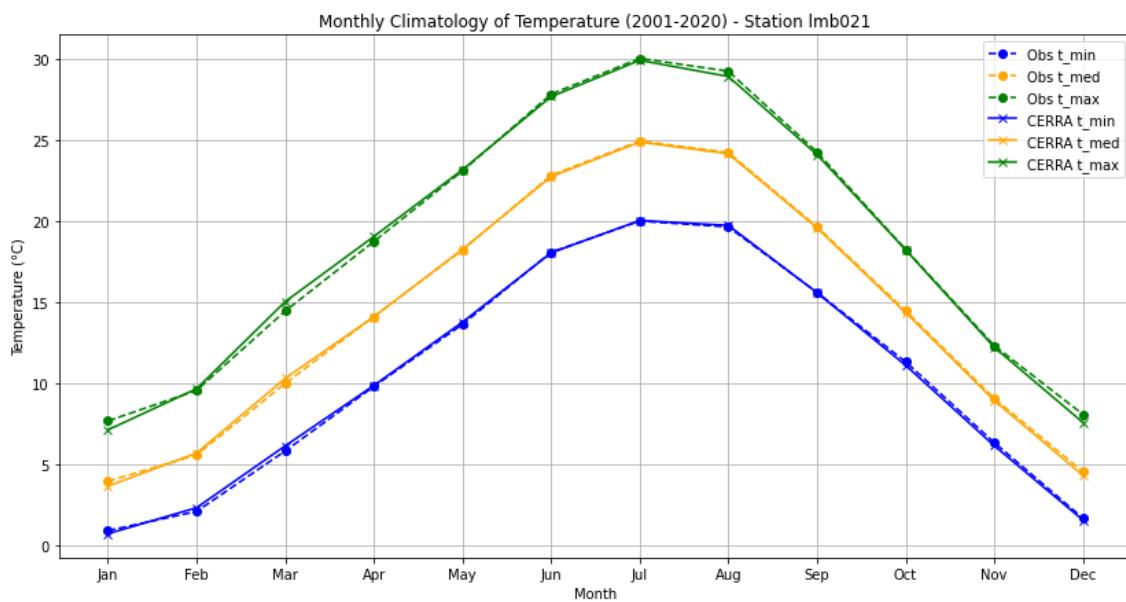


Figure 24: Comparison between monthly reconstructed climatologies and the ones obtained from observation during the common period, for station *lmb021*.

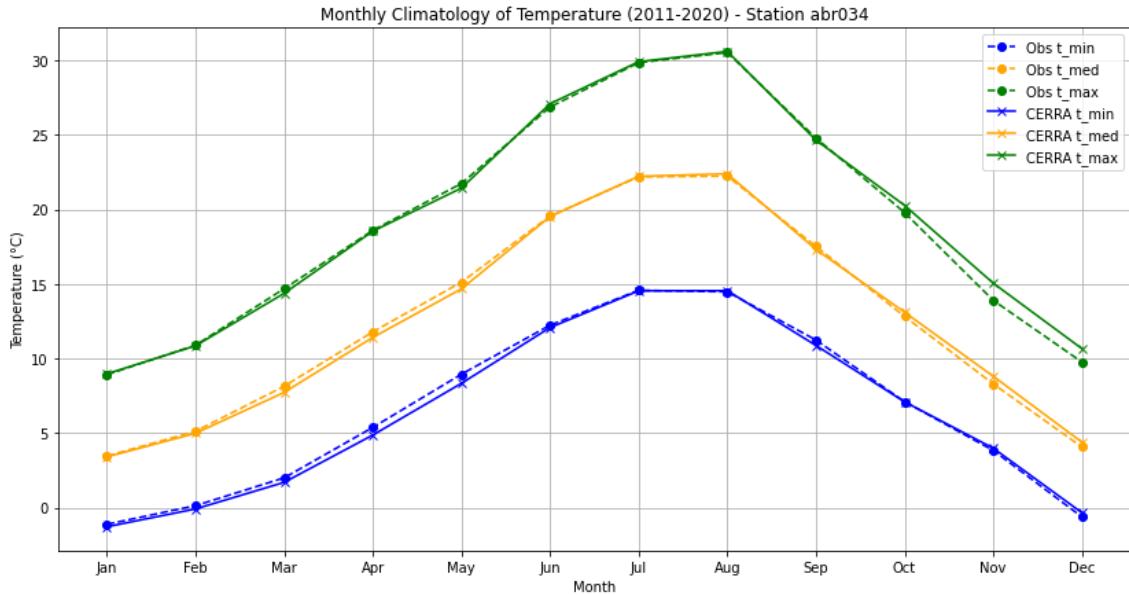


Figure 25: Comparison between monthly reconstructed climatologies and the ones obtained from observation during the common period, for station abr034.

In contrast, the selected stations for precipitation are as follows: the most accurate is *tsc001*, while the least accurate is *ero001*, shown in Figure 26 and 27, respectively.

With regard to the Tuscany station, it can be observed that the cumulative monthly precipitation averages exhibit a relatively limited discrepancy from the observed values. The greatest divergence is observed in April, with a difference of approximately 7.5 mm (11.5%). The remaining months consistently demonstrate a discrepancy of less than 6 mm, with an average deviation of 5.8%. With regard to the mountainous station of Emilia Romagna, a discrepancy of approximately 20 mm (39%) is observed between the two climatologies for the initial six-month period. It is therefore evident that the model is unsuitable for accurately representing the precipitation climate trend of the station.

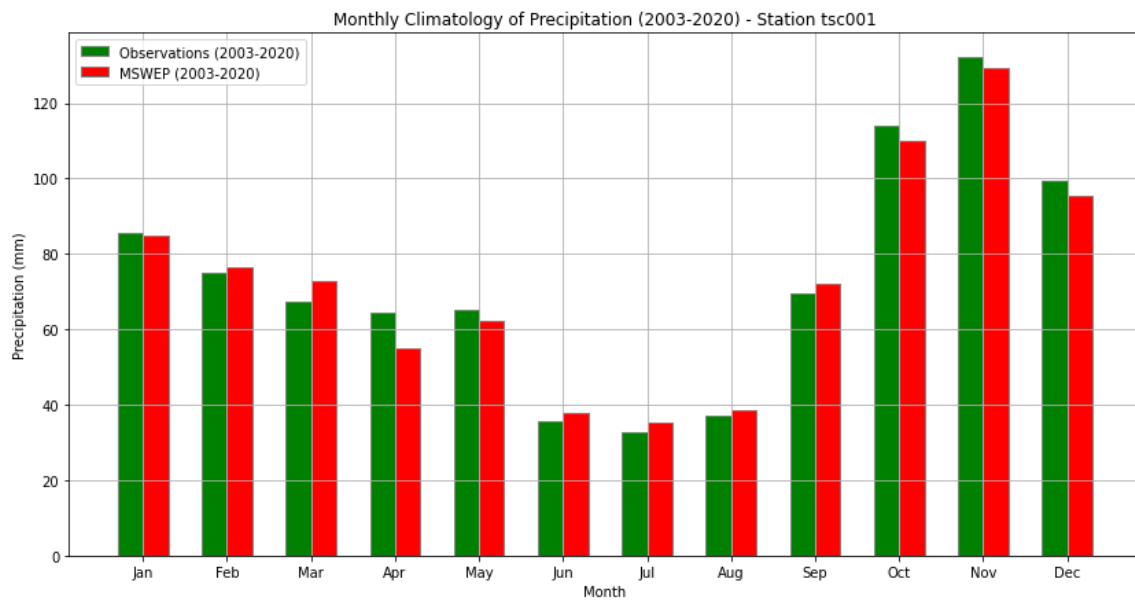


Figure 26: Comparison between monthly climatologies of precipitation reconstructed and the ones obtained from the observation on the common period for station tsc001.

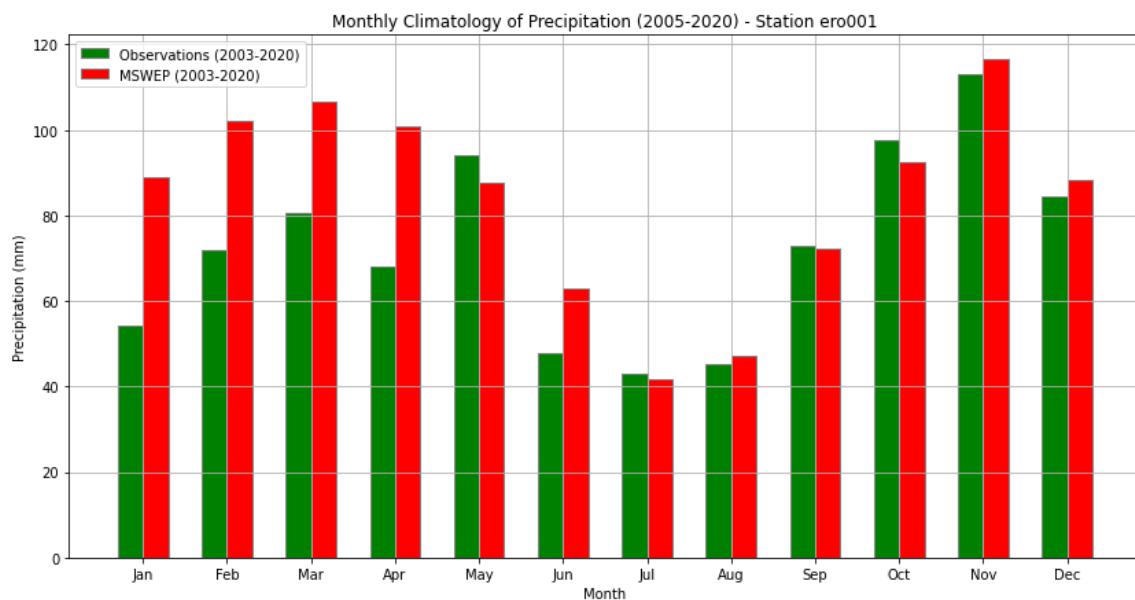


Figure 27: Comparison between monthly climatologies of precipitation reconstructed and the ones obtained from the observation on the common period for station ero001.



## 4 Conclusion

In this study was developed a methodology for reconstructing climatologies for the period 1991-2020 using daily data of maximum, minimum, and average temperatures, as well as cumulative precipitation, from the meteorological stations of the Citizen Science network Meteonetwork.

In order to identify an effective methodology, 43 stations with the longest time series of the variables of interest were selected from the MNW database. The historical time series for each variable of interest of the selected stations range from 9 to 23 years. It is therefore necessary to identify a method for reconstructing the time series of the stations in order to ensure the availability of at least 30 years of observations for the calculation of climatologies.

The methodology was based on the use of the closest grid points obtained from a gridded reanalysis model, with the objective of establishing a relationship between reanalysis data and station data over a specified period (reference) through machine learning (ML) techniques. Once the relationship has been identified, it was used to reconstruct an equivalent synthetic time series for each station over the climatological period 1991-2020, which was then employed to calculate the climatological normals.

A number of reanalysis datasets were selected for testing, in order to ascertain which one should be chosen for input into the ML procedure. Five different reanalysis datasets were analysed for temperature, two of which were global and three regional. The global datasets were Era5-Land and MSWX, while the regional ones were CERRA, MERIDA\_H\_RES and VHR\_REA\_IT. In addition to the aforementioned reanalyses, two further precipitation datasets were considered: MSWEP and CHIRPS.

Two types of machine learning (ML) models were also selected, which have previously demonstrated efficacy in similar contexts: Support Vector Regression (SVR) and Random Forest (RF).

In order to obtain a model that reconstructs the MNW station time series as accurately as possible, it is essential to select a suitable input, and therefore it is essential to select a reanalysis that has a high degree of compatibility with the observations recorded at all stations. To this end, each reanalysis dataset is compared with all the observations from all 43 stations together, first by comparing the raw reanalysis data and then by comparing the reanalysis data corrected with simple procedures that take into account the distance of the grid points from the station, by interpolation with IDW, and the differences in altitude between the grid points and the stations. These corrections were made because global and regional reanalyses often show systematic and regionally distributed biases with respect to observations from meteorological ground stations.

For **temperature**, the reanalysis data set with the best agreement with observations for all temperature variables was **CERRA**. It was therefore decided to continue the analysis using this dataset. In addition, Era5-Land was chosen as a complementary analysis as it has large temperature time series extending to mid-2023, a period two years longer than the time series available for CERRA. The aim was to test whether this could be a factor influencing the performance of the machine learning models. **ERA5-Land** is also the “runner-up” reanalysis after CERRA, when elevation correction is applied.

In order to assess the models' performance in simulating **precipitation**, the following three models were selected: ERA5-Land, CERRA and MSWEP. The **MSWEP** model demonstrated the most optimal performance, exhibiting the highest correlation coefficient and the lowest RMSE. The **CERRA** and **ERA5-Land** models, which exhibited a relatively close standard deviation to the observed data, were selected as the second and third best models, respectively.

The results demonstrate that the optimal performance is not necessarily achieved with higher-resolution reanalyses or those downscaled specifically for the Italian region. Conversely, global reanalyses with lower resolution demonstrate superior performance in this instance. With regard to temperature, the European-scale reanalysis proved superior, whereas in the case of precipitation, the performance of even two global reanalyses was more favourable.

Subsequently, for each variable and station, all potential combinations between the distinct model types (SVR and RF) and the selected reanalysis datasets (CERRA, ERA5-Land, MSWEP) were evaluated to determine the optimal performance, with varying values of the models' hyper-parameters. This was achieved through a Grid Search procedure to optimise the parameter values and k-fold cross-validation to prevent overfitting of the models.

With regard to temperature, all the tested models demonstrated high accuracy, with  $R^2$  values exceeding 0.85 and RMSE values consistently below 2°C in validation for all the stations and the variables. The models exhibited an enhanced aptitude for representing **mean temperature**, achieving the highest  $R^2$  values (between 0.96 and 0.99) and the lowest RMSEs, which were equal to or less than 1.5°C. In contrast, the reconstruction of **minimum temperature** exhibited the poorest performance, with  $R^2$  values between 0.85 and 0.98 and RMSE values between 1 and 2.5°C. On the other hand, **maximum temperature** demonstrated a more favourable outcome, with  $R^2$  values between 0.90 and 0.98 and RMSE values between 1 and 2°C. The two models, Support Vector Regression (SVR) and Random Forest (RF), exhibited comparable and highly analogous performance when the optimal parameters for each station were identified. However, **SVR with linear kernel** demonstrated slight superiority in the majority of cases. The latter exhibited optimal performance in 67% of instances for maximum temperature, 81% for mean temperature and 84% for minimum temperature.

Similarly, the two reanalyses (CERRA and ERA5-Land) demonstrated comparable performance when employed as inputs to the same model, with minimal discrepancy. The only notable discrepancy was observed at the single station situated at an altitude exceeding 2,000 metres (*trn017*). The reconstruction of the minimum temperature using the ERA5-Land reanalysis yielded an  $R^2$  of 0.86 and an RMSE of 2.5°C, whereas with CERRA, the  $R^2$  was 0.94 and the RMSE was 1.5°C. This may indicate that the higher resolution of CERRA (5.5 km) in comparison to ERA5-Land (9 km) in high-elevation mountainous regions contributes to enhanced performance.

The models employed for the reconstruction of **precipitation** series were considerably more complex than those utilised for temperature series. The incorporation of **sample weights** into the models proved to be a crucial factor in enhancing their predictive capacity. In the absence of such weights, the models were unable to identify an optimal function, resulting in the generation of negative  $R^2$  values in some cases. The optimal weight value, which yields the best performance, is  $10^4$  for days with precipitation and zero for days without precipitation. The values of the **coefficient of determination** demonstrate considerable variability in the process of validation, with a range of 0.1 to 0.8 depending on the specific station under consideration. In the case of 11 stations (26% of the total), it was not possible to obtain a model that exceeded the performance of  $R^2$  equal to 0.5 in the validation process. In contrast, the **RMSEs** demonstrate a relatively consistent error, varying between approximately **3 and 6 millimetres**. In contrast to temperature, which exhibited minimal variability, the efficacy of precipitation models is contingent upon both the specific model employed and the reanalysis data utilised as input.

The only model that demonstrated an acceptable level of accuracy for all observation stations was the **SVR with a radial (RBF) kernel**. In contrast, the use of SVR with a linear kernel required an excessive amount of computational time for the Grid Search method to identify optimal hyper-parameters, rendering the process impractical on a standard laptop within a limited timeframe. While Random Forests demonstrated

favourable outcomes at certain stations, they exhibited inferior performance compared to SVR and were effective only at a very restricted number of stations. The most effective reanalysis for precipitation was **MSWEP**, which demonstrated optimal performance in 77% of cases.

In order to reconstruct the temperature and precipitation series and calculate the climatologies in a unified manner, it was decided that input reanalyses and models which demonstrated superior performance on average across the entire sample of stations should be selected. Accordingly, **the temperature series were reconstructed using support vector regression (SVR) with a linear kernel and the CERRA reanalysis, whereas the precipitation series were reconstructed using SVR with a radial kernel and the MSWEP reanalysis**, based on the average performance obtained at all stations.

Finally, in order to verify the accuracy of the reconstructed climatologies, a comparison was conducted between these and the climatologies calculated using only the daily observations of the available stations, despite the presence of missing data, during the period of overlapping. Although this comparison is not an exact reflection of the actual data, as the calibration and validation of the models occurred during the same period, it provides a means of determining, in addition to the model accuracy indicators, whether the reconstructed series are sufficiently similar to the actual series.

With regard to temperature, it can be observed that even under the most unfavourable circumstances, the monthly climatologies show a minimum deviation of a few degrees from the actual data, thereby demonstrating their effectiveness as a representation of climatological averages. It can, therefore, be reasonably assumed that the results will be favourable even in periods where there is no overlap, due to the good accuracy of the obtained models.

With regard to precipitation, a comparison of the best station climatologies reveals that the cumulative monthly averages of precipitation exhibit a relatively low degree of

divergence from the actual recorded values. The greatest divergence is observed in April, with a difference of approximately 7.5 mm (11.5%). The remaining months always exhibit a discrepancy of less than 6 mm, with an average deviation of 5.8%. It can be also observed that the model underestimates precipitation in half of the months, particularly during the winter months, while overestimating it in the summer months.

For the station exhibiting the poorest model performance, there is a discrepancy of about 20 mm (39%) between the climatologies for the initial six-month period. It is therefore evident that the model is unsuitable for accurately representing the climate trend of the station.

It was thus demonstrated that the designed methodology was highly effective in the reconstruction of temperature series, thereby yielding reliable climatologies in instances where complete historical data were unavailable. However, the performance for precipitation was more variable, with the accuracy of the model dependent on the station and the input data. It seems probable that this discrepancy can be attributed to issues pertaining to the measurement of observed station data. An alternative explanation may be found in the absence of heated rain gauges at numerous stations. Furthermore, even the regional reanalysis datasets may be unable to resolve localised phenomena such as thunderstorms, which is likely to have contributed to the poorer performance of precipitation models. The MSWEP dataset had a resolution of about 10 km, which was insufficient for a number of stations. This highlights the need for more detailed input data in future analyses. Moreover, precipitation exhibits markedly disparate temporal and spatial gradients when compared to temperature. This renders the training of an ML model for precipitation more challenging. In the absence of the sample weights, the model is unable to discern the relationship between the reanalysis data and observations.

In conclusion, this approach represents a promising solution to the problem of incomplete observational data, enabling the reconstruction of climatological time

series. Nevertheless, the efficacy of this methodology is contingent upon the quality of both the station observations and the reanalysis data employed. In particular, further advancements in reanalysis resolution and model capabilities will be essential to improving the accuracy of precipitation reconstruction in areas where localised weather phenomena play a significant role.



## Bibliography

- [1] K. Calvin *et al.*, «IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.», Intergovernmental Panel on Climate Change (IPCC), lug. 2023. doi: 10.59327/IPCC/AR6-9789291691647.
- [2] M.-K. Lee, S.-H. Moon, Y. Yoon, Y.-H. Kim, e B.-R. Moon, «Detecting Anomalies in Meteorological Data Using Support Vector Regression», *Adv. Meteorol.*, vol. 2018, pp. 1–14, giu. 2018, doi: 10.1155/2018/5439256.
- [3] C. Daly, W. Gibson, G. Taylor, G. Johnson, e P. Pasteris, «A knowledge-based approach to the statistical mapping of climate», *Clim. Res.*, vol. 22, pp. 99–113, 2002, doi: 10.3354/cr022099.
- [4] M. C. Llasat *et al.*, «High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database», *Adv. Geosci.*, vol. 23, pp. 47–55, mar. 2010, doi: 10.5194/adgeo-23-47-2010.
- [5] M. M. Miglietta e S. Davolio, «Dynamical forcings in heavy precipitation events over Italy: lessons from the HyMeX SOP1 campaign», *Hydrol. Earth Syst. Sci.*, vol. 26, fasc. 3, pp. 627–646, feb. 2022, doi: 10.5194/hess-26-627-2022.
- [6] M. Adinolfi, M. Raffa, A. Reder, e P. Mercogliano, «Investigation on potential and limitations of ERA5 Reanalysis downscaled on Italy by a convection-permitting model», *Clim. Dyn.*, vol. 61, fasc. 9–10, pp. 4319–4342, nov. 2023, doi: 10.1007/s00382-023-06803-w.
- [7] M. Brunetti, M. Maugeri, T. Nanni, C. Simolo, e J. Spinoni, «High-resolution temperature climatology for Italy: interpolation method intercomparison», *Int. J. Climatol.*, vol. 34, fasc. 4, pp. 1278–1296, mar. 2014, doi: 10.1002/joc.3764.
- [8] A. Crespi, M. Brunetti, G. Lentini, e M. Maugeri, «1961–1990 high-resolution monthly precipitation climatologies for Italy», *Int. J. Climatol.*, vol. 38, fasc. 2, pp. 878–895, feb. 2018, doi: 10.1002/joc.5217.
- [9] C. Daly *et al.*, «Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States», *Int. J. Climatol.*, vol. 28, fasc. 15, pp. 2031–2064, dic. 2008, doi: 10.1002/joc.1688.

- [10] M. Giazzì *et al.*, «Meteonetwork: An Open Crowdsourced Weather Data System», *Atmosphere*, vol. 13, fasc. 6, p. 928, giu. 2022, doi: 10.3390/atmos13060928.
- [11] U. Wehn *et al.*, «Impact assessment of citizen science: state of the art and guiding principles for a consolidated approach», *Sustain. Sci.*, vol. 16, fasc. 5, pp. 1683–1699, set. 2021, doi: 10.1007/s11625-021-00959-2.
- [12] G. R. C. Essou, F. Sabarly, P. Lucas-Picher, F. Brissette, e A. Poulin, «Can Precipitation and Temperature from Meteorological Reanalyses Be Used for Hydrological Modeling?», *J. Hydrometeorol.*, vol. 17, fasc. 7, pp. 1929–1950, lug. 2016, doi: 10.1175/JHM-D-15-0138.1.
- [13] C. Soci, E. Bazile, F. Besson, e T. Landelius, «High-resolution precipitation re-analysis system for climatological purposes», *Tellus Dyn. Meteorol. Oceanogr.*, vol. 68, fasc. 1, p. 29879, dic. 2016, doi: 10.3402/tellusa.v68.29879.
- [14] M. Ridal *et al.*, «CERRA , the Copernicus European Regional Reanalysis system», *Q. J. R. Meteorol. Soc.*, p. qj.4764, mag. 2024, doi: 10.1002/qj.4764.
- [15] A. Pelosi, F. Terribile, G. D'Urso, e G. Chirico, «Comparison of ERA5-Land and UERRA MESCAN-SURFEX Reanalysis Data with Spatially Interpolated Weather Observations for the Regional Assessment of Reference Evapotranspiration», *Water*, vol. 12, fasc. 6, p. 1669, giu. 2020, doi: 10.3390/w12061669.
- [16] C. Shen, «A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists», *Water Resour. Res.*, vol. 54, fasc. 11, pp. 8558–8593, nov. 2018, doi: 10.1029/2018WR022643.
- [17] O. M. Baez-Villanueva *et al.*, «RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements», *Remote Sens. Environ.*, vol. 239, p. 111606, mar. 2020, doi: 10.1016/j.rse.2019.111606.
- [18] A. Kumar, R. Ramsankaran, L. Brocca, e F. Munoz-Arriola, «A Machine Learning Approach for Improving Near-Real-Time Satellite-Based Rainfall Estimates by Integrating Soil Moisture», *Remote Sens.*, vol. 11, fasc. 19, p. 2221, set. 2019, doi: 10.3390/rs11192221.
- [19] G. V. Nguyen, X.-H. Le, L. Nguyen Van, D. T. T. May, S. Jung, e G. Lee, «Machine learning approaches for reconstructing gridded precipitation based on multiple source products», *J. Hydrol. Reg. Stud.*, vol. 48, p. 101475, ago. 2023, doi: 10.1016/j.ejrh.2023.101475.
- [20] H. Hersbach *et al.*, «The ERA5 global reanalysis», *Q. J. R. Meteorol. Soc.*, vol. 146, fasc. 730, pp. 1999–2049, lug. 2020, doi: 10.1002/qj.3803.

- [21] Copernicus Climate Change Service, «ERA5-Land hourly data from 1950 to present». Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2019. doi: 10.24381/CDS.E2161BAC.
- [22] Copernicus Climate Change Service, «CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present». ECMWF, 2021. doi: 10.24381/CDS.622A565A.
- [23] R. Bonanno, M. Lacavalla, e S. Sperati, «A new high-resolution Meteorological Reanalysis Italian Dataset: MERIDA», *Q. J. R. Meteorol. Soc.*, vol. 145, fasc. 721, pp. 1756–1779, apr. 2019, doi: 10.1002/qj.3530.
- [24] H. E. Beck *et al.*, «MSWX: Global 3-Hourly 0.1° Bias-Corrected Meteorological Data Including Near-Real-Time Updates and Forecast Ensembles», *Bull. Am. Meteorol. Soc.*, vol. 103, fasc. 3, pp. E710–E732, mar. 2022, doi: 10.1175/BAMS-D-21-0145.1.
- [25] M. Raffa *et al.*, «VHR-REA\_IT Dataset: Very High Resolution Dynamical Downscaling of ERA5 Reanalysis over Italy by COSMO-CLM», *Data*, vol. 6, fasc. 8, p. 88, ago. 2021, doi: 10.3390/data6080088.
- [26] H. E. Beck *et al.*, «MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment», *Bull. Am. Meteorol. Soc.*, vol. 100, fasc. 3, pp. 473–500, mar. 2019, doi: 10.1175/BAMS-D-17-0138.1.
- [27] C. Funk *et al.*, «The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes», *Sci. Data*, vol. 2, fasc. 1, p. 150066, dic. 2015, doi: 10.1038/sdata.2015.66.
- [28] A. J. Smola e B. S. Scholkopf, «A tutorial on support vector regression», *Stat. Comput.*, vol. 14, fasc. 3, pp. 199–222, 2004.
- [29] S.-T. Chen e P.-S. Yu, «Pruning of support vector networks on flood forecasting», *J. Hydrol.*, vol. 347, fasc. 1–2, pp. 67–78, 2007, doi: 10.1016/j.jhydrol.2007.08.029.
- [30] S. Tripathi, V. V. Srinivas, e R. S. Nanjundiah, «Downscaling of precipitation for climate change scenarios: A support vector machine approach», *J. Hydrol.*, vol. 330, fasc. 3–4, pp. 621–640, 2006, doi: 10.1016/j.jhydrol.2006.04.030.
- [31] L. Zhang *et al.*, «Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach», *J. Hydrol.*, vol. 594, 2021, doi: 10.1016/j.jhydrol.2021.125969.
- [32] G. Biau e E. Scornet, «A random forest guided tour», *Test*, vol. 25, fasc. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

- [33] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, fasc. 1, pp. 5–32, 2001, doi: <https://doi.org/10.1023/A:1010933404324>.
- [34] K. E. Taylor, «Summarizing multiple aspects of model performance in a single diagram», *J. Geophys. Res. Atmospheres*, vol. 106, fasc. D7, pp. 7183–7192, apr. 2001, doi: 10.1029/2000JD900719.
- [35] J. Muñoz-Sabater *et al.*, «ERA5-Land: a state-of-the-art global reanalysis dataset for land applications», *Earth Syst. Sci. Data*, vol. 13, fasc. 9, pp. 4349–4383, set. 2021, doi: 10.5194/essd-13-4349-2021.

# Appendix

The following appendix presents the graphical representations of the climatologies of all 43 selected stations, as resulting from the calculations conducted within the framework of this thesis. It deserves noting that the temperature climatologies of all stations are considered reliable; however, this is not the case for precipitation climatologies, which are only deemed reliable if the station model has demonstrated sufficient satisfactory performance.

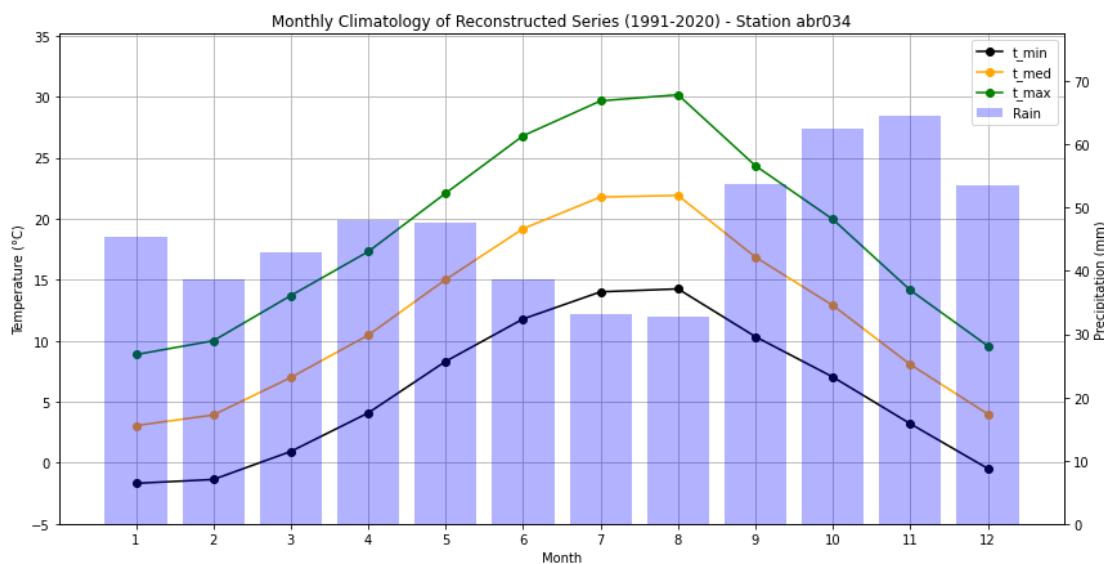
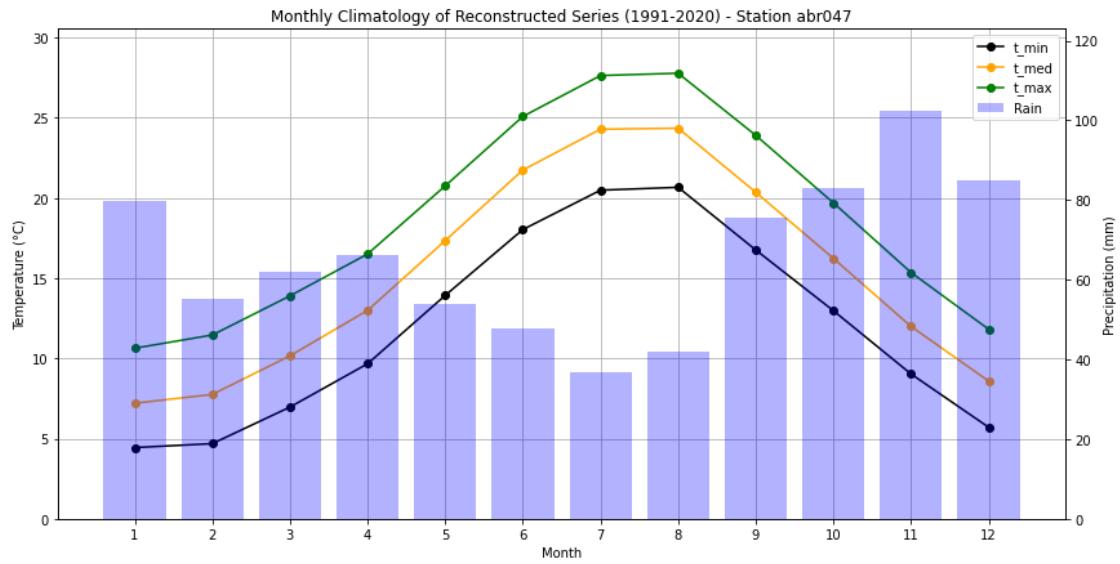


Figure A.1: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station abr034 during the period 1991-2020.



FigureA.2: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station abr047 during the period 1991-2020.

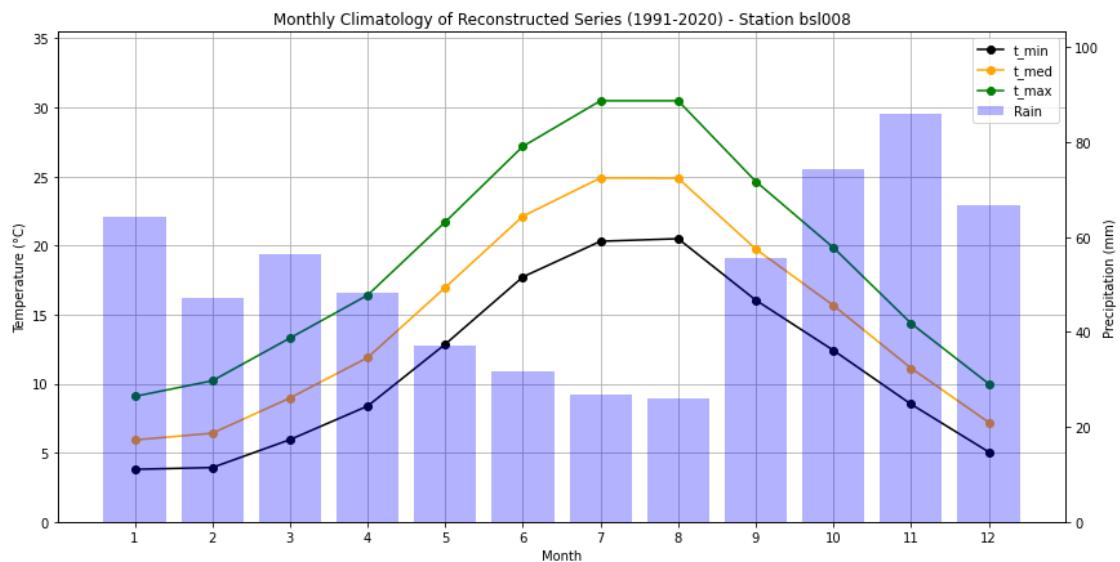


Figure A.3: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station bsl008 during the period 1991-2020.

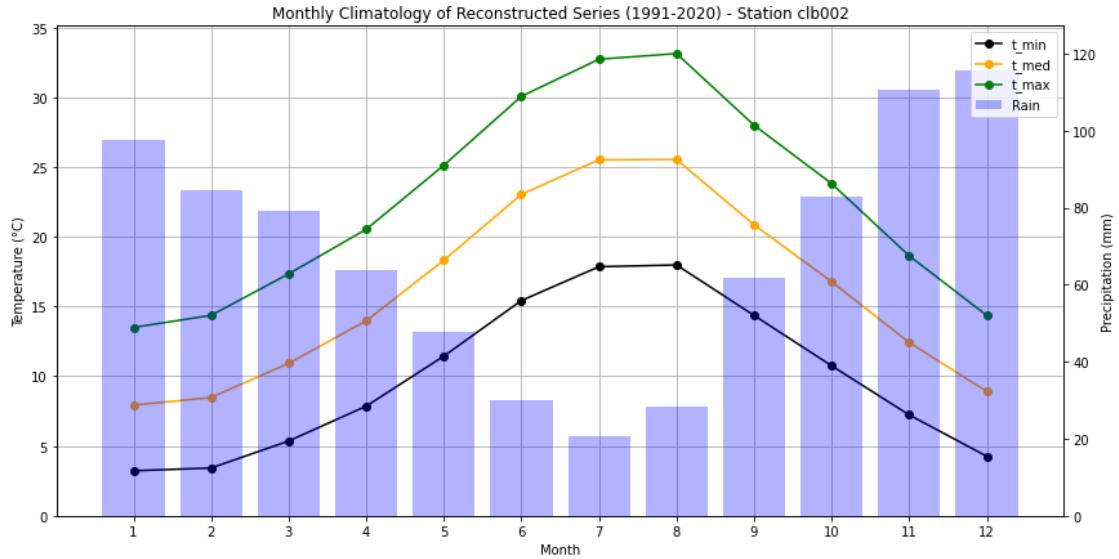


Figure A.4: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station clb002 during the period 1991-2020.

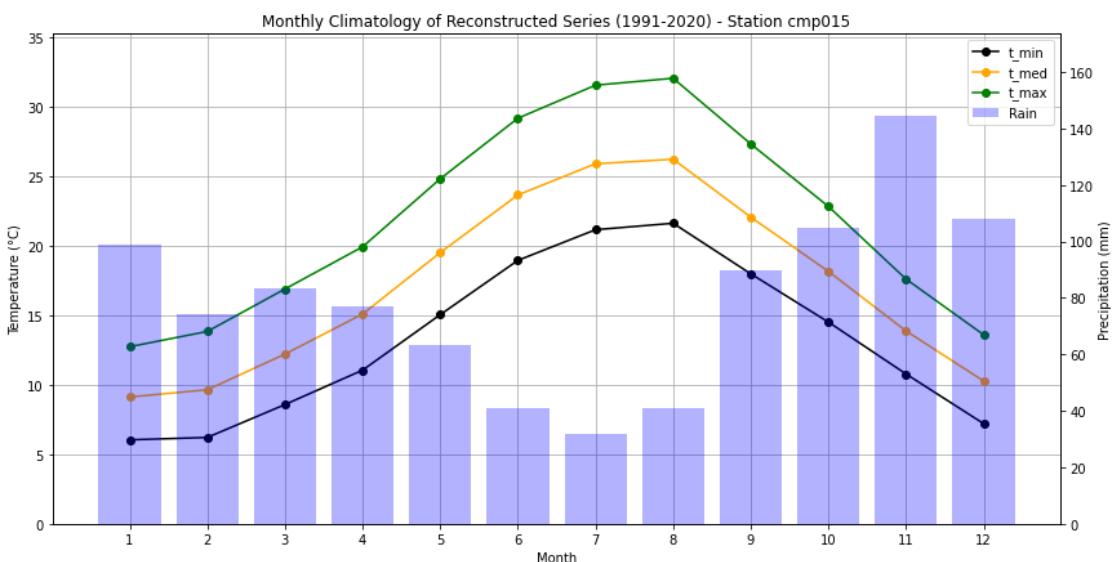


Figure A.5: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station cmp015 during the period 1991-2020.

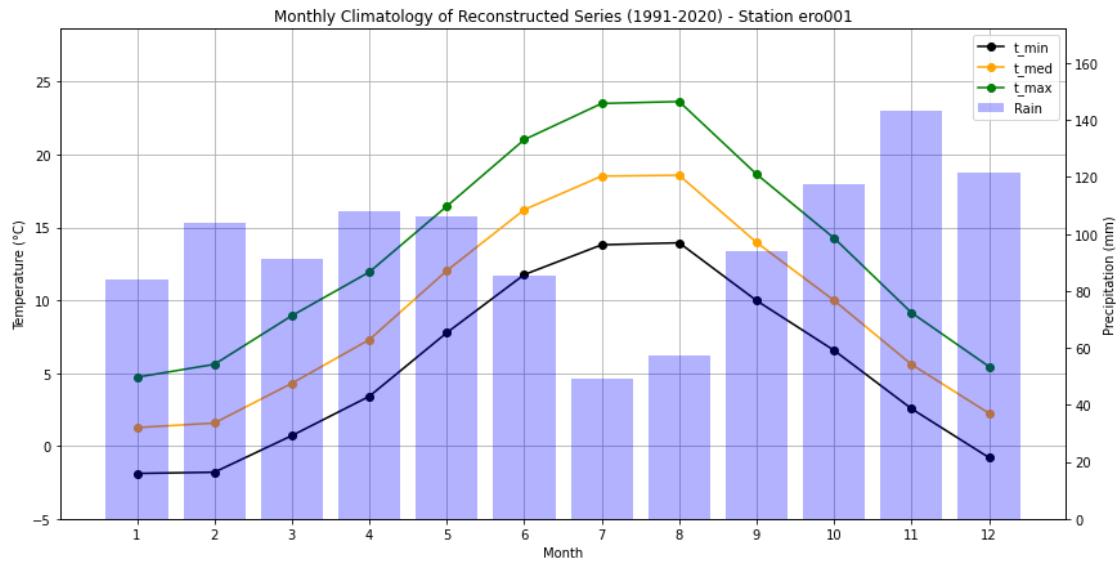


Figure A.6: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station ero001 during the period 1991-2020.

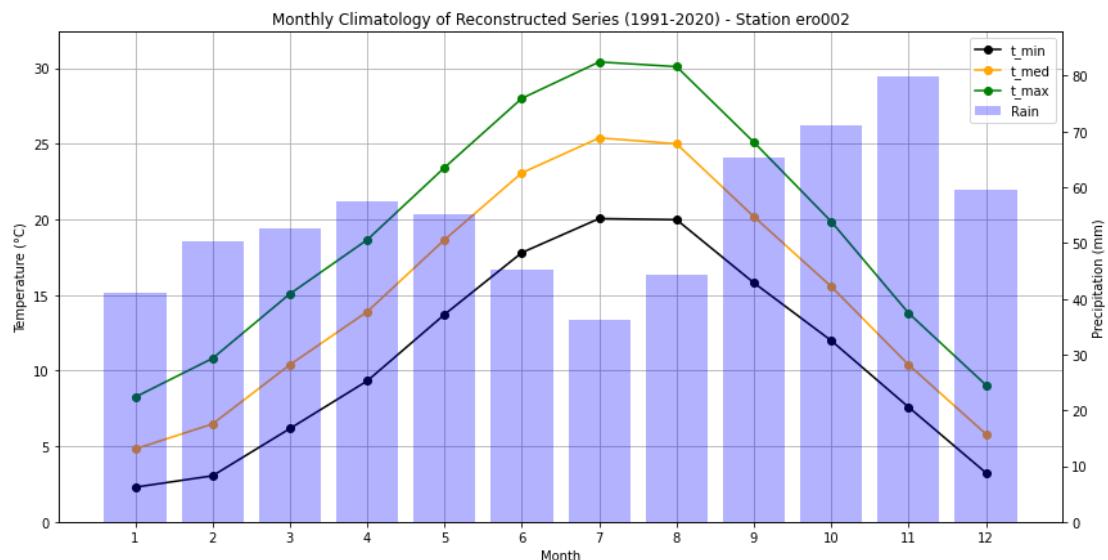


Figure A.7: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station ero002 during the period 1991-2020.

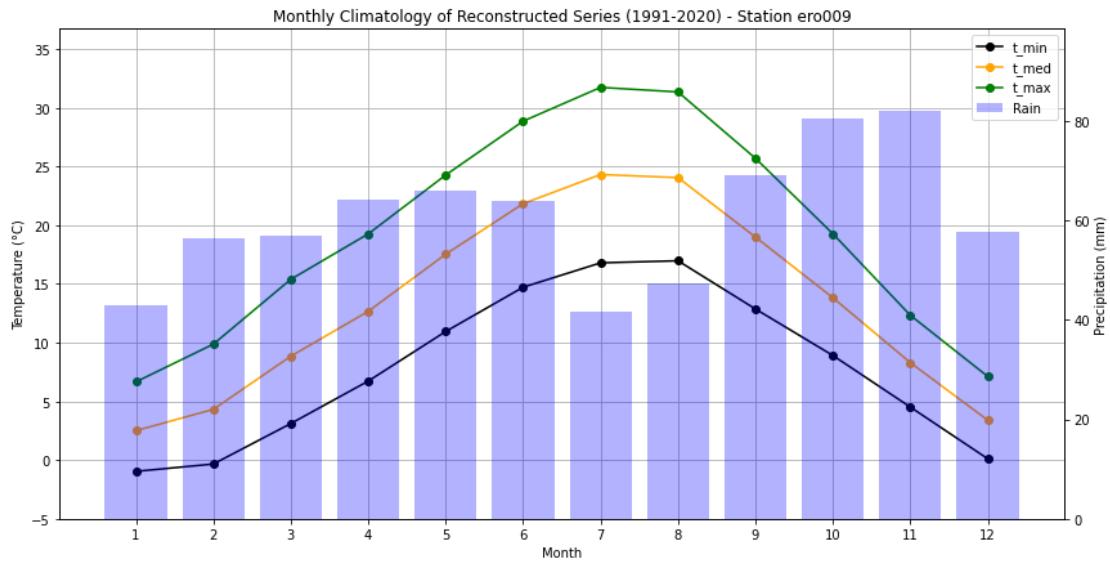


Figure A.8: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station ero009 during the period 1991-2020.

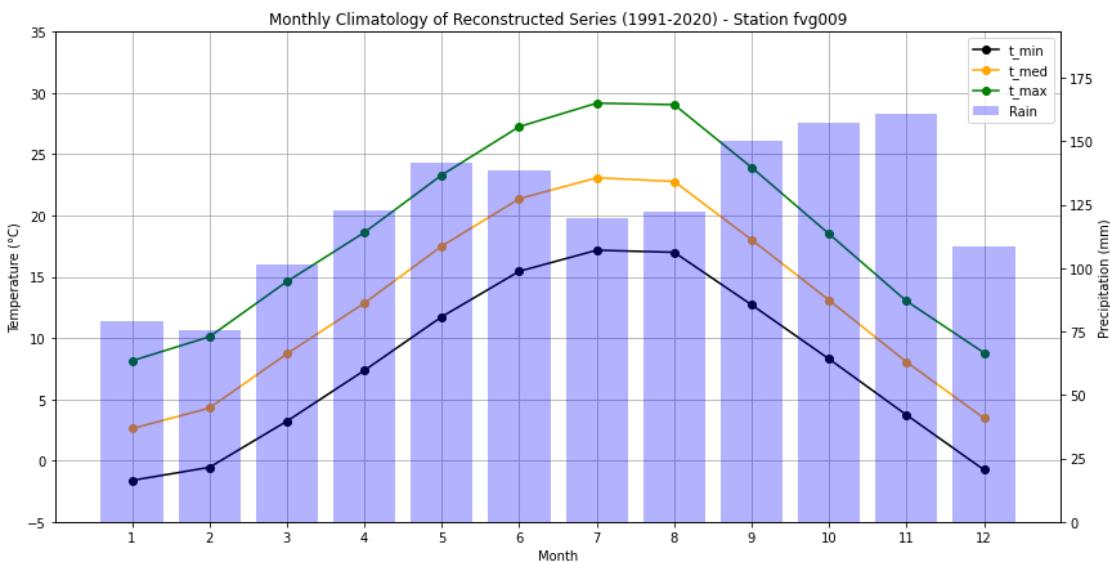


Figure A.9: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station fvg009 during the period 1991-2020.

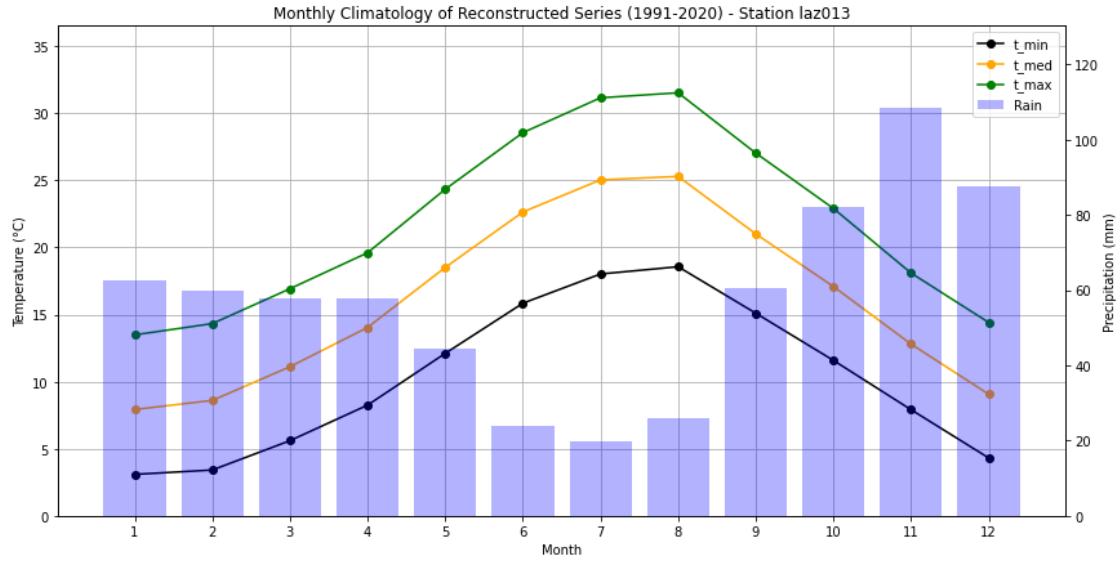


Figure A.10: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station laz013 during the period 1991-2020.

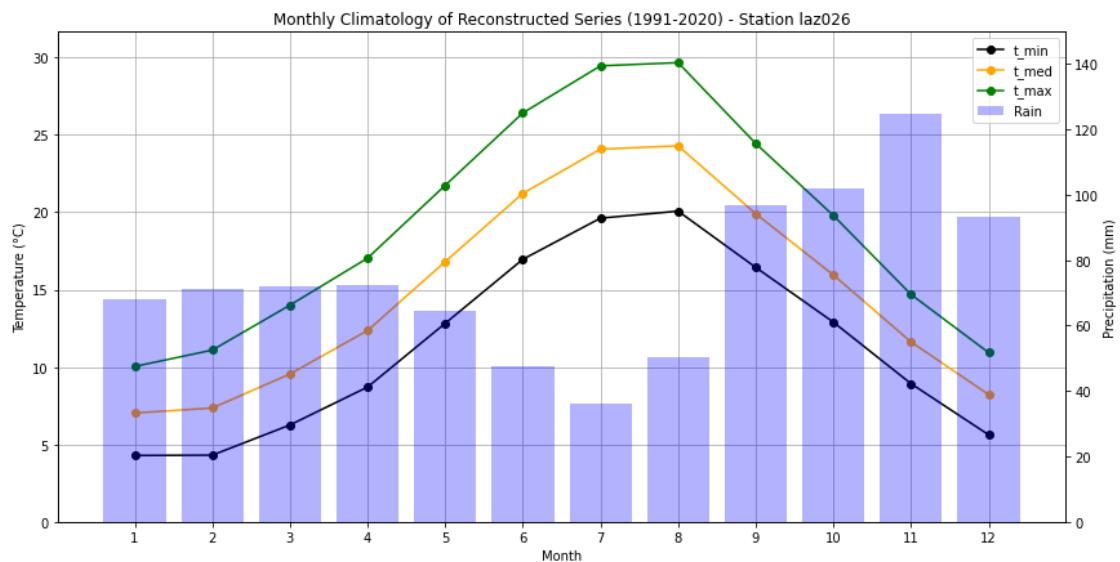


Figure A.11: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station laz026 during the period 1991-2020.

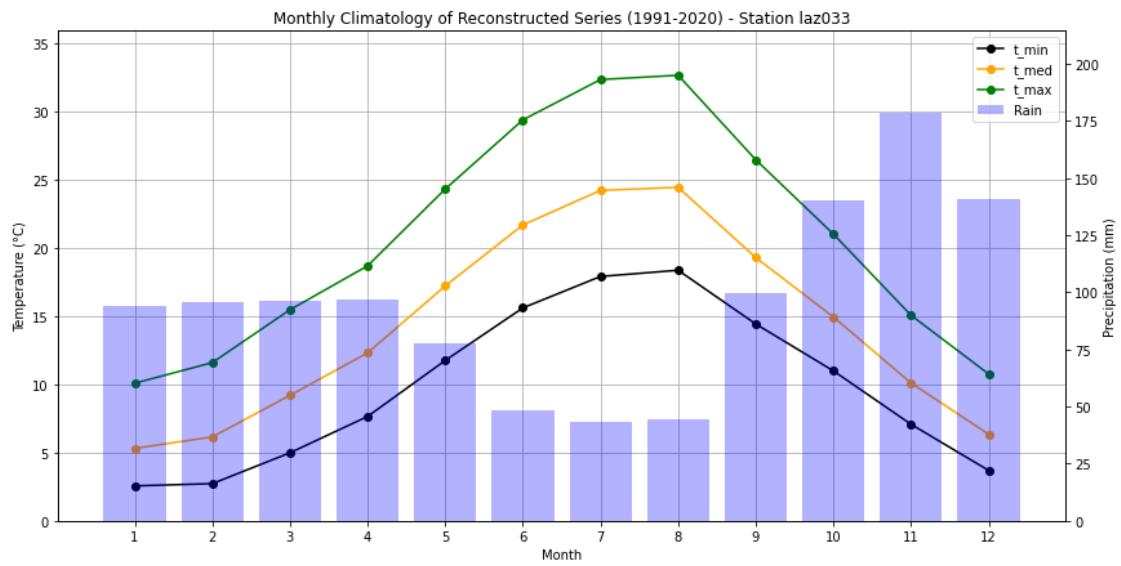


Figure A.12: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station laz033 during the period 1991-2020.

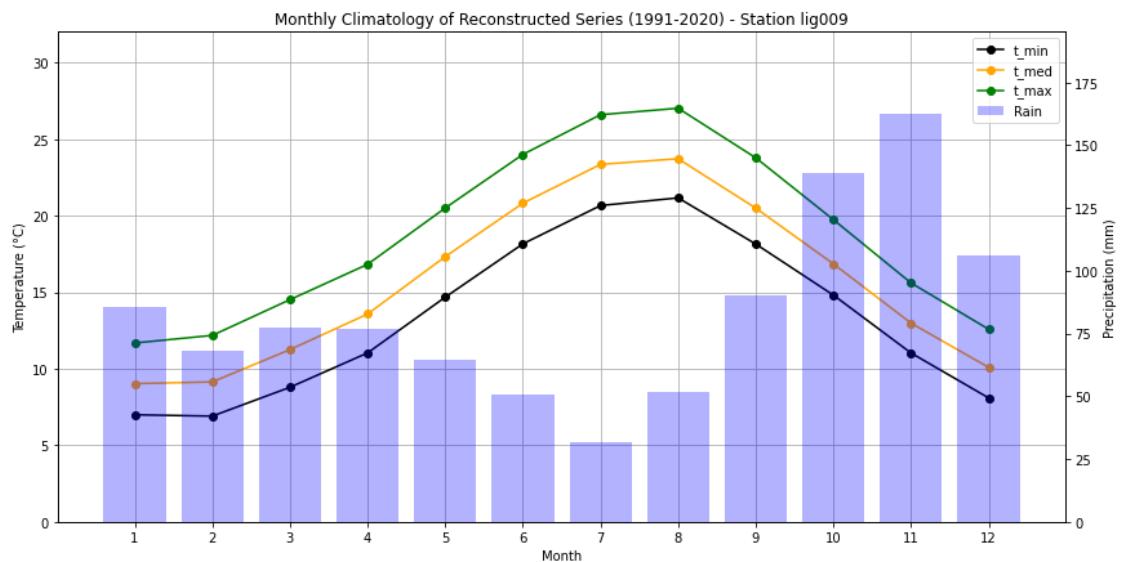


Figure A.13: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lig009 during the period 1991-2020.

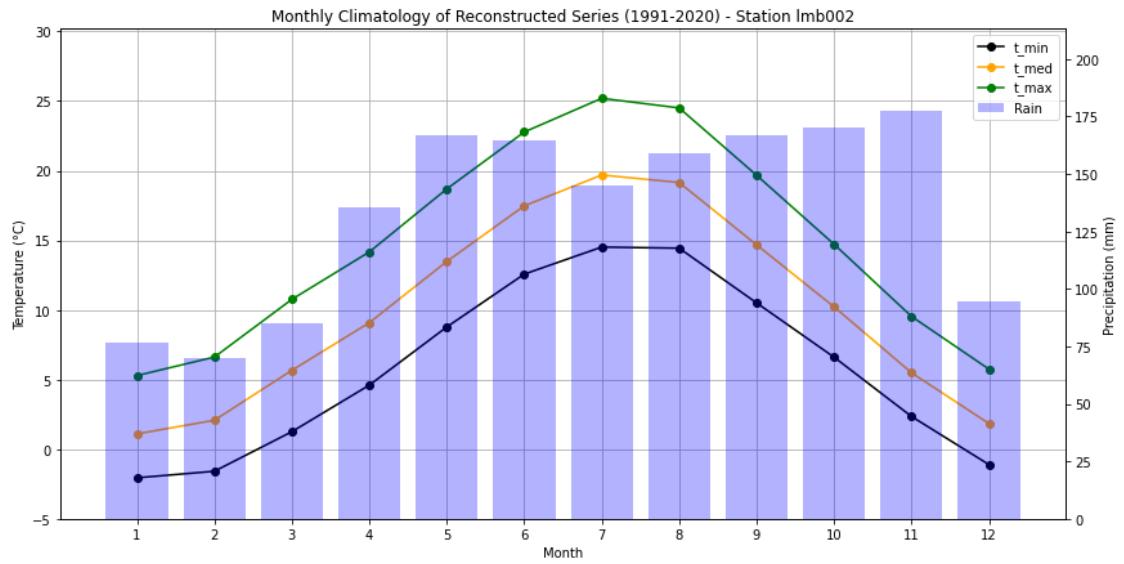


Figure A.14: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb002 during the period 1991-2020.

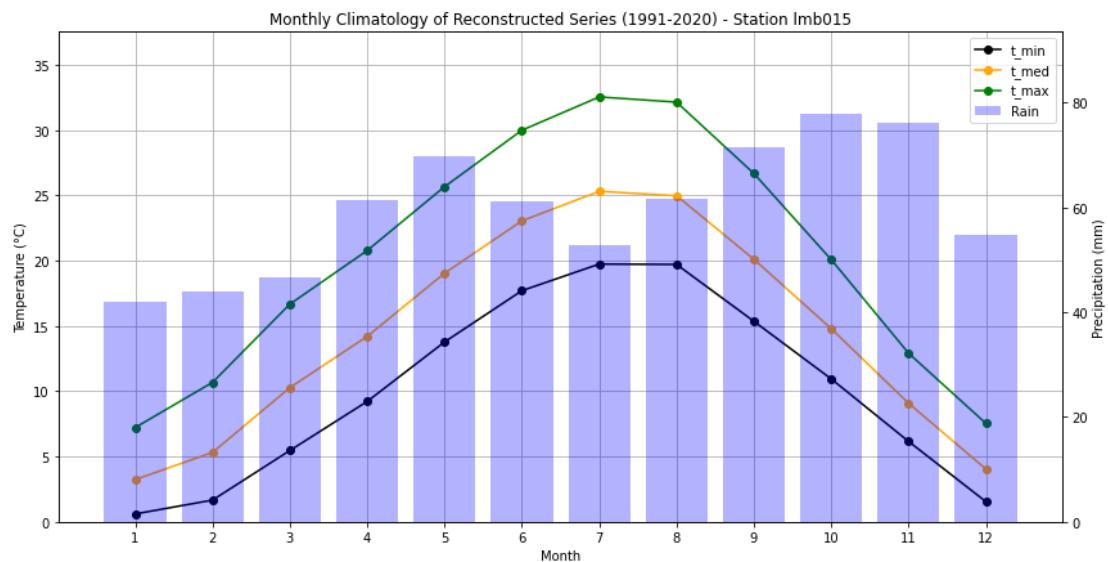


Figure A.15: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb015 during the period 1991-2020.

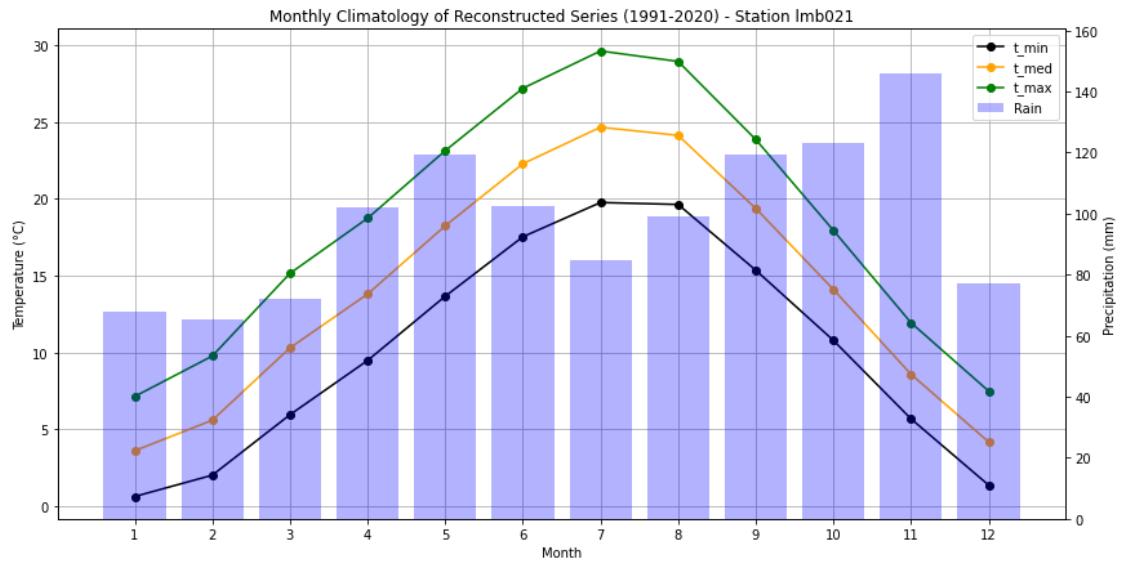


Figure A.16: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb021 during the period 1991-2020.

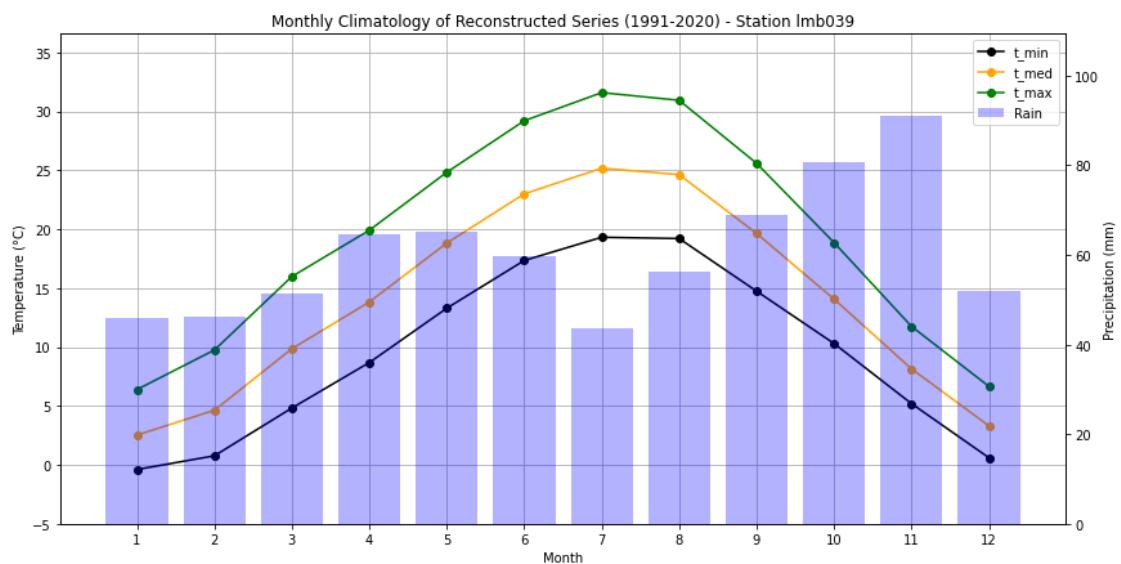


Figure A.17: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb039 during the period 1991-2020.

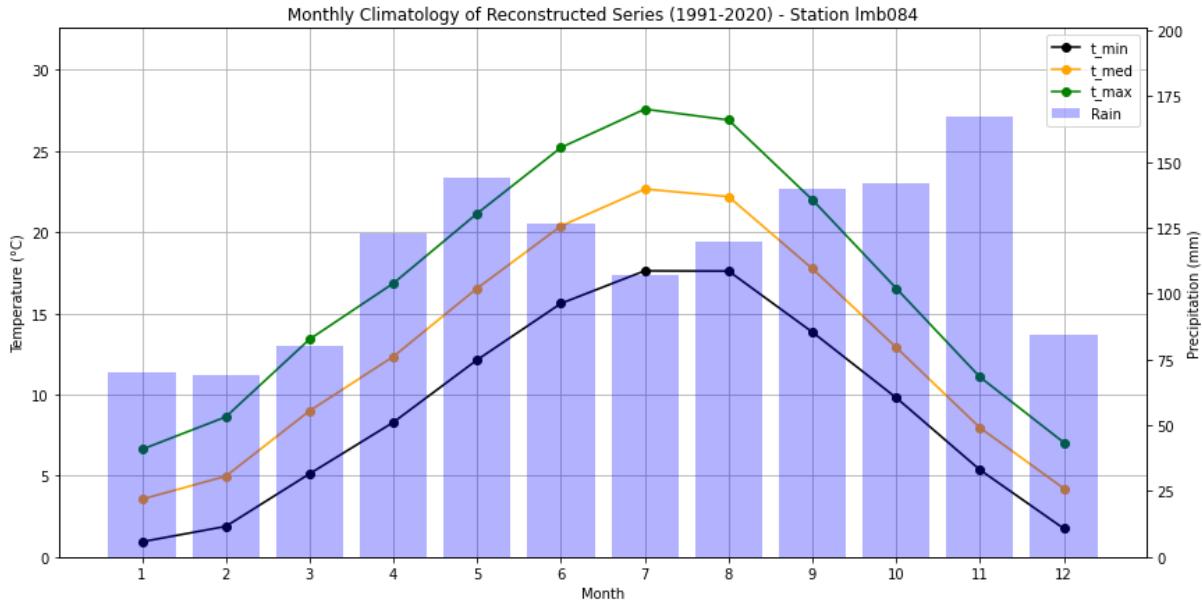


Figure A.18: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb084 during the period 1991-2020.

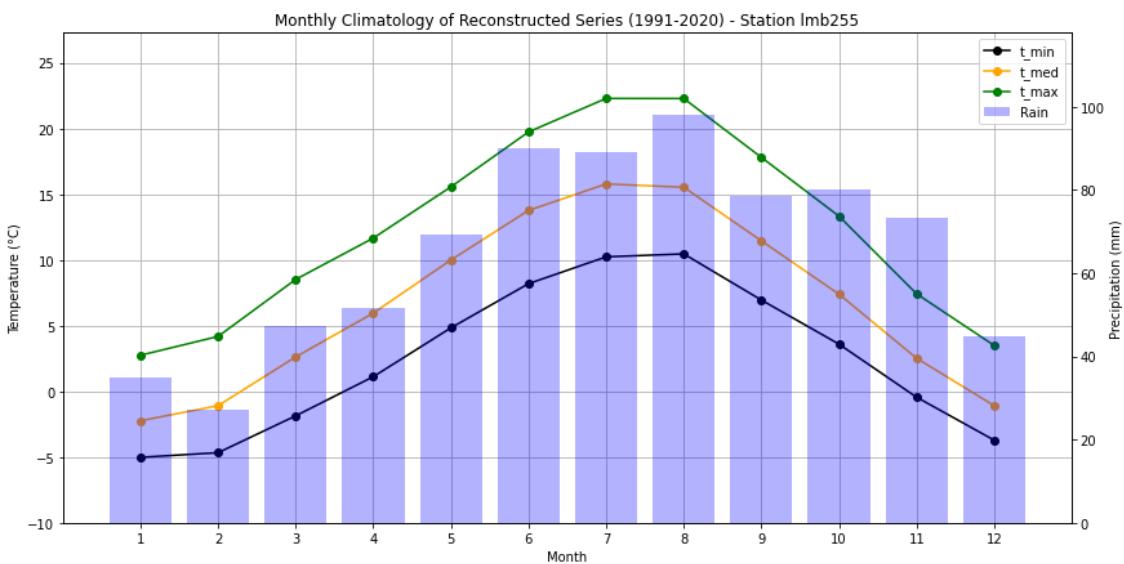


Figure A.19: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station lmb255 during the period 1991-2020.

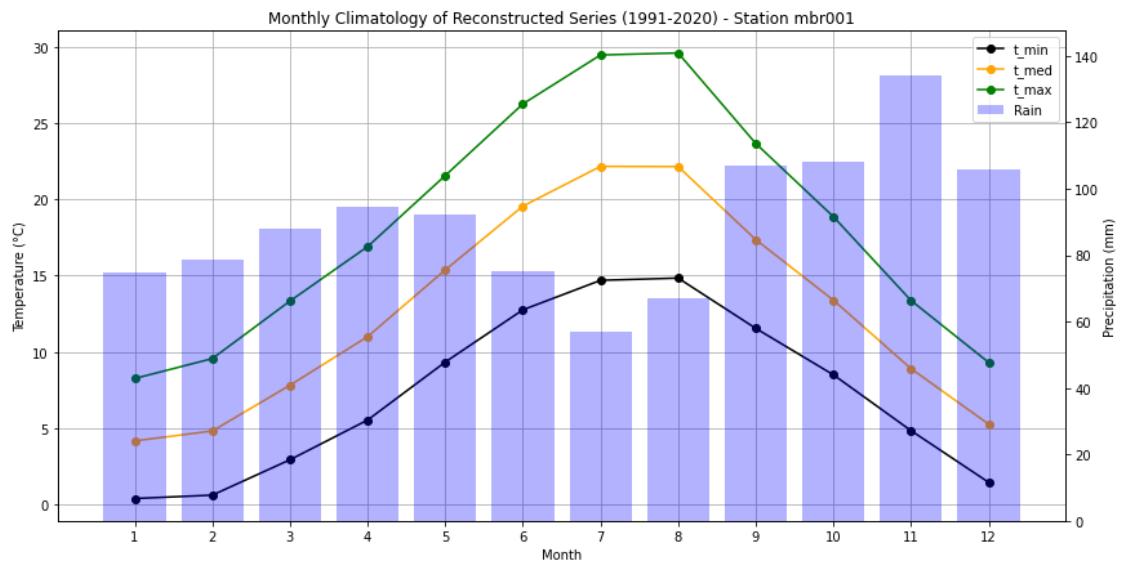


Figure A.20: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station mbr001 during the period 1991-2020.

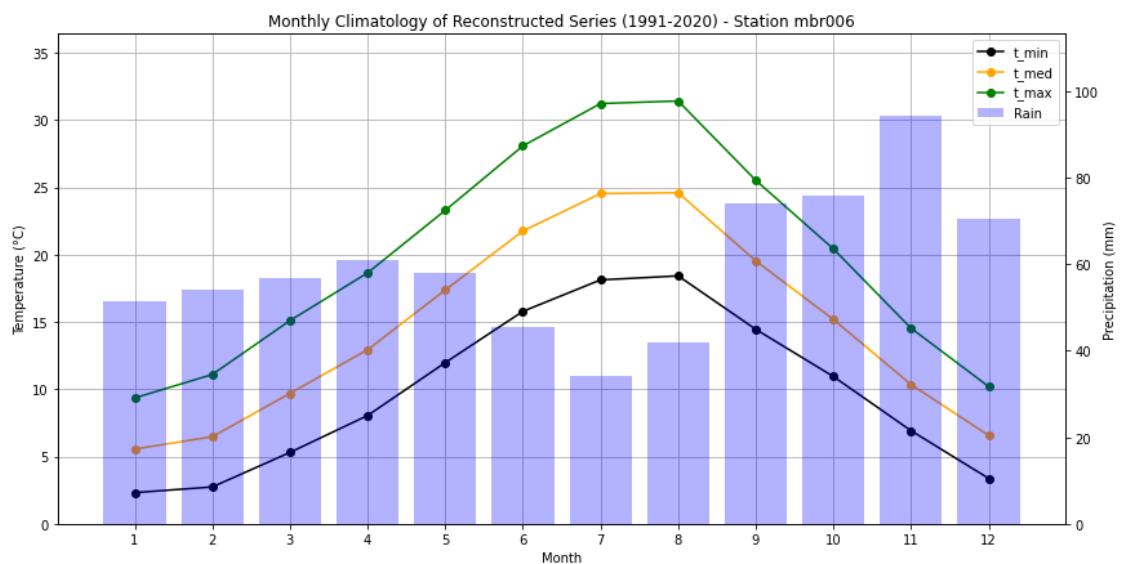


Figure A.21: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station mbr006 during the period 1991-2020.

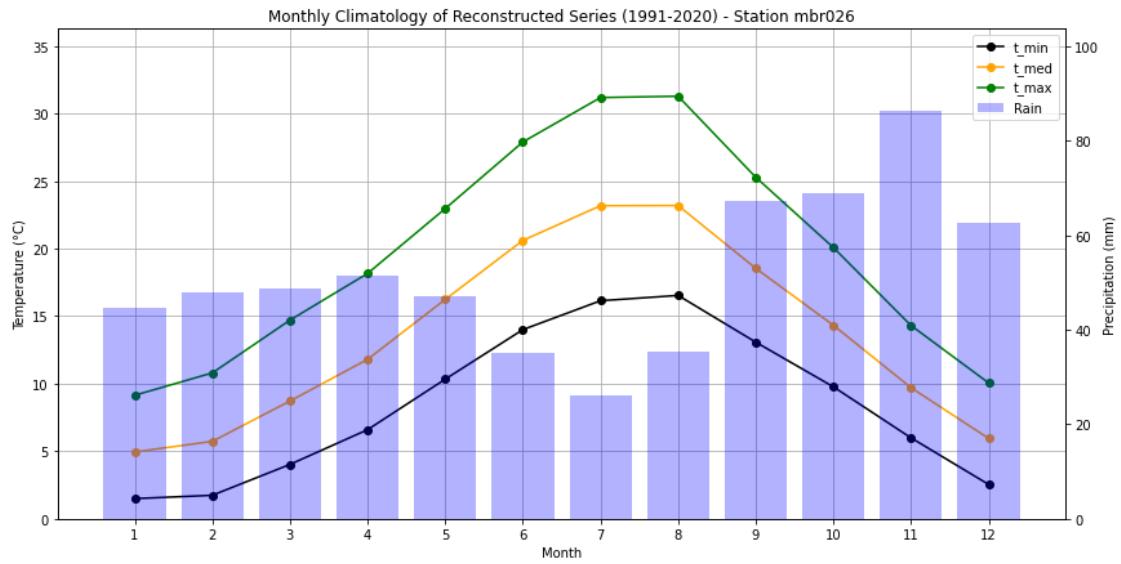


Figure A.22: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station mbr026 during the period 1991-2020.

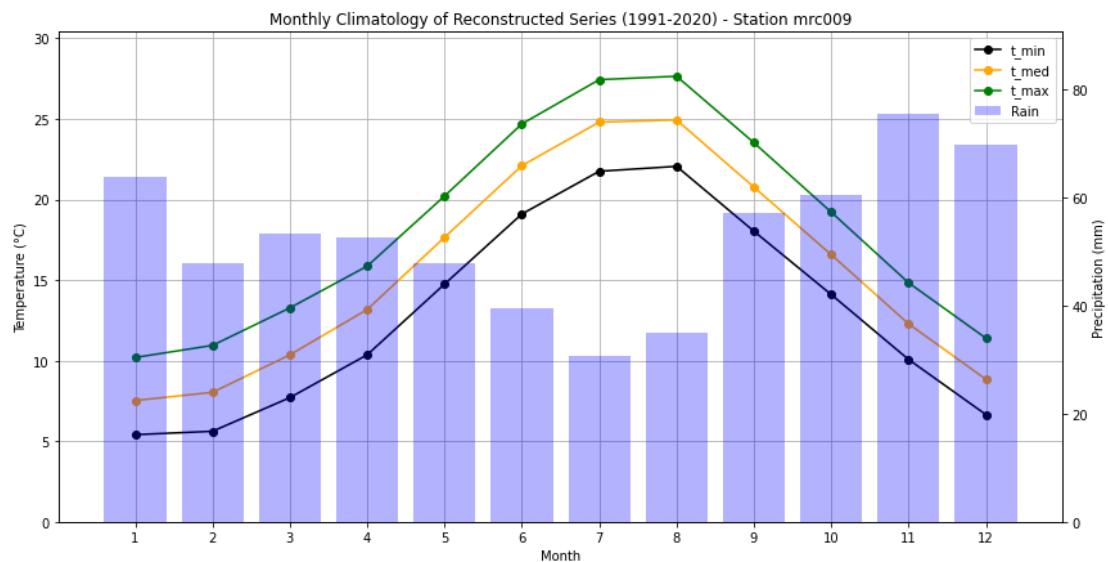


Figure A.23: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station mrc009 during the period 1991-2020.

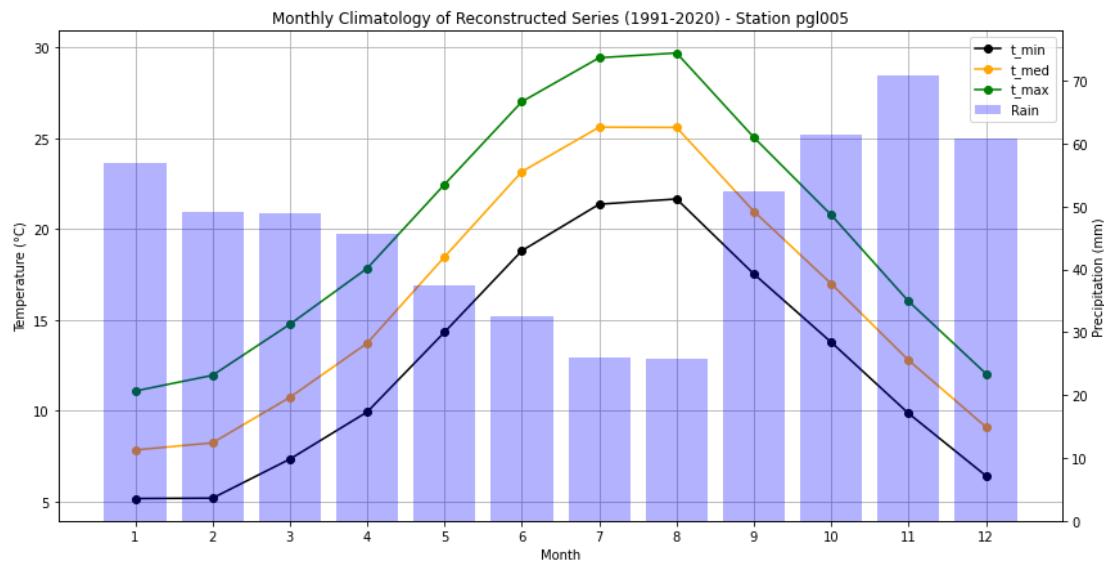


Figure A.24: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pgl005 during the period 1991-2020.

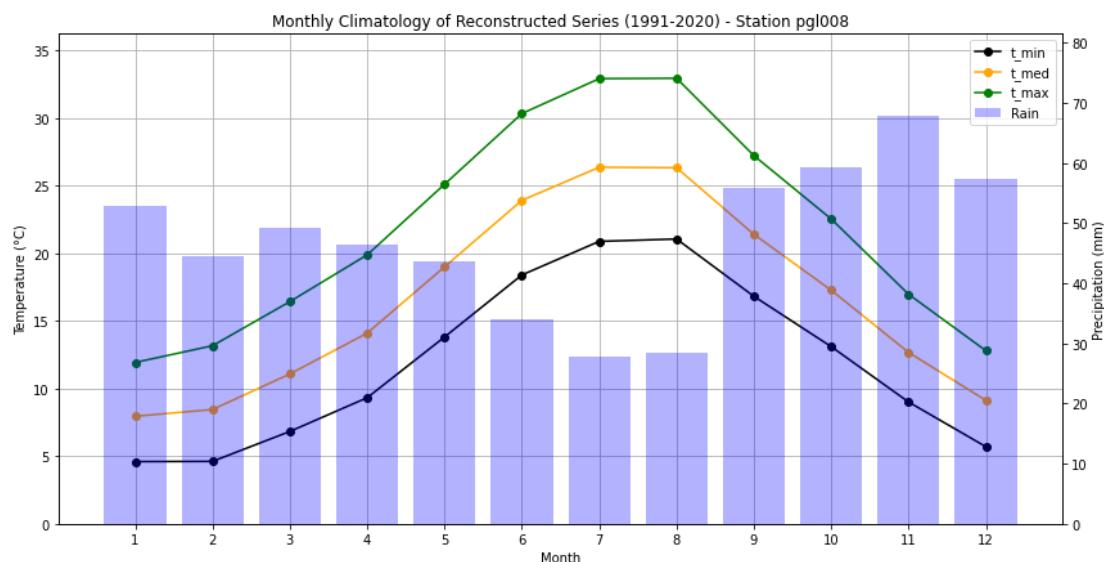


Figure A.25: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pgl008 during the period 1991-2020.

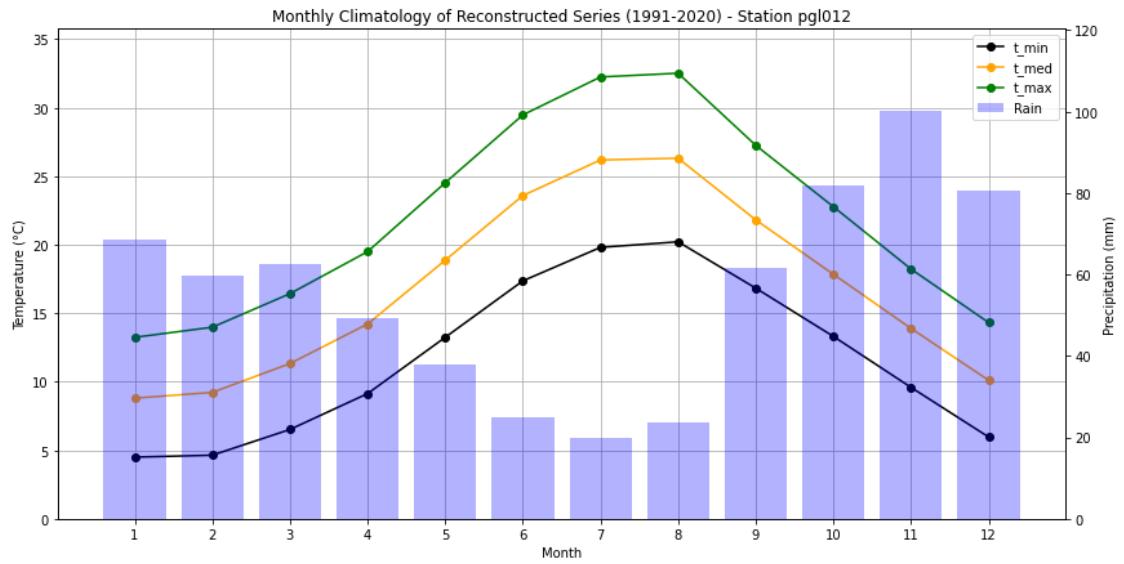


Figure A.26: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pgl012 during the period 1991-2020.

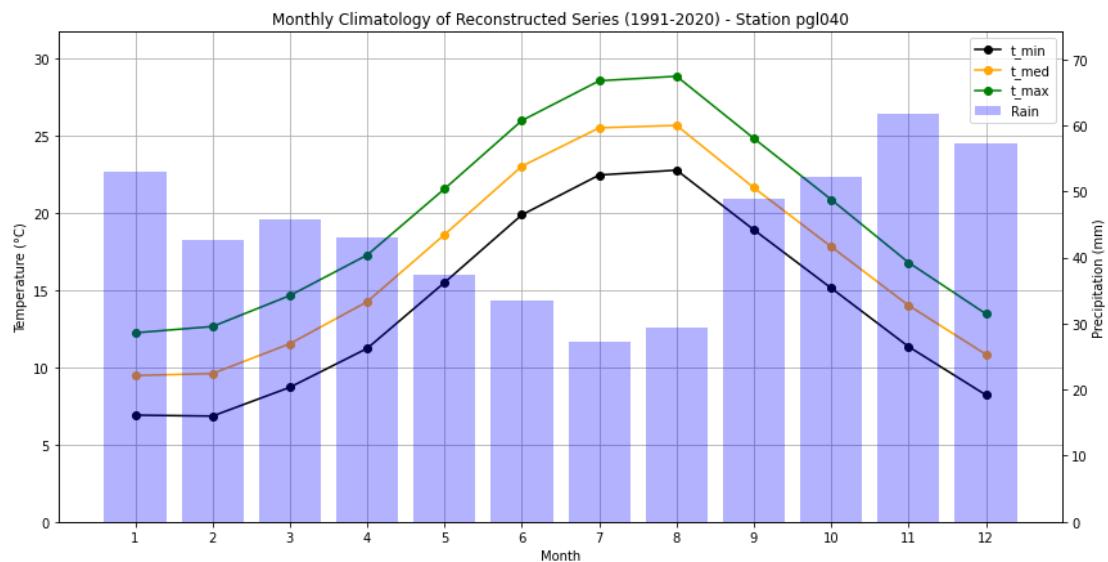


Figure A.27: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pgl040 during the period 1991-2020.

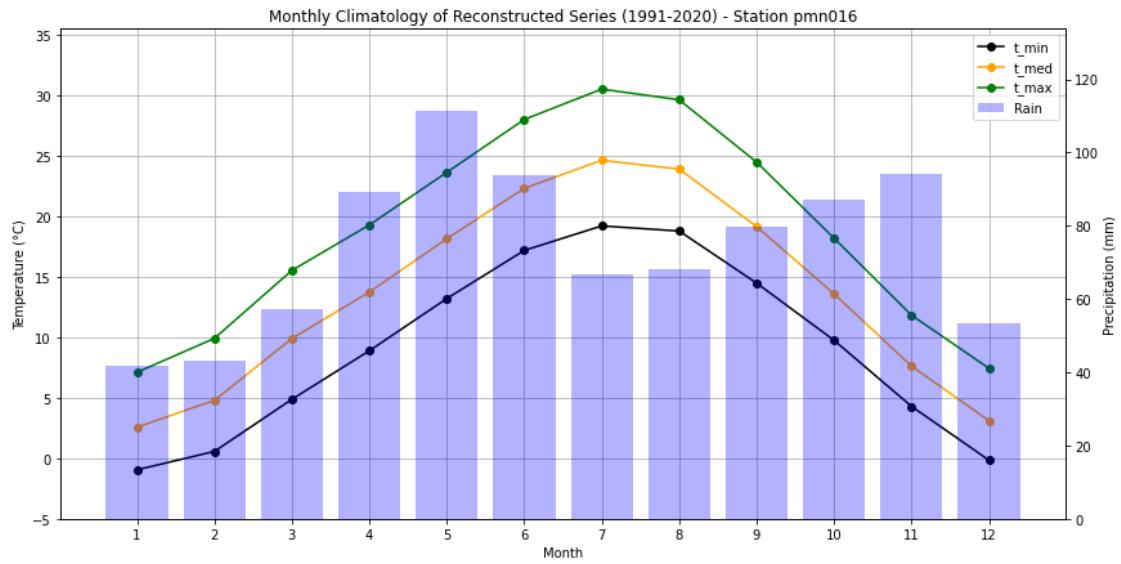


Figure A.28: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pmn016 during the period 1991-2020.

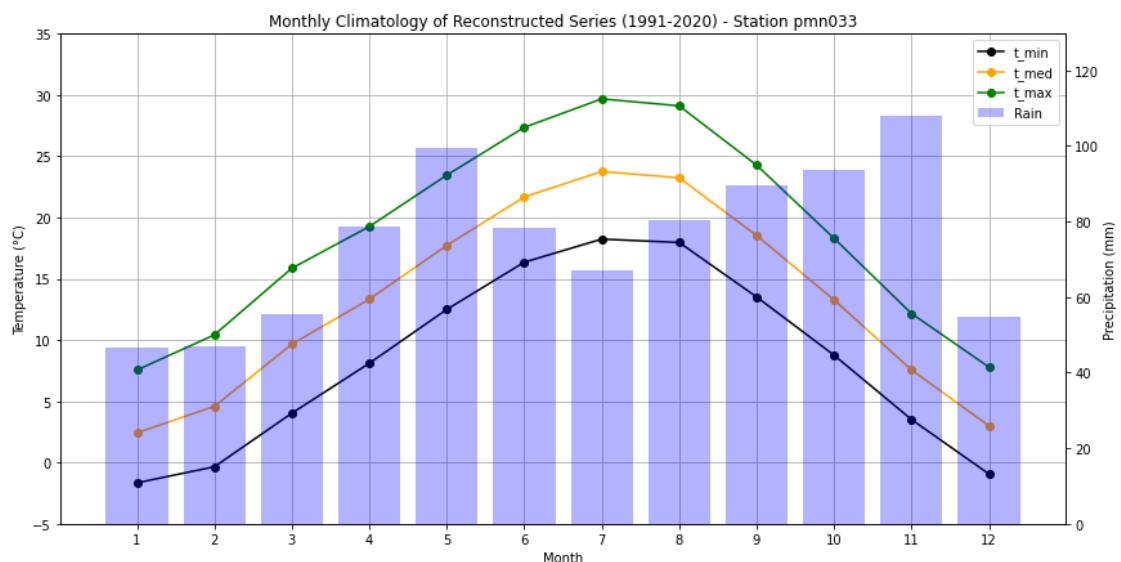


Figure A.29: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pmn033 during the period 1991-2020.

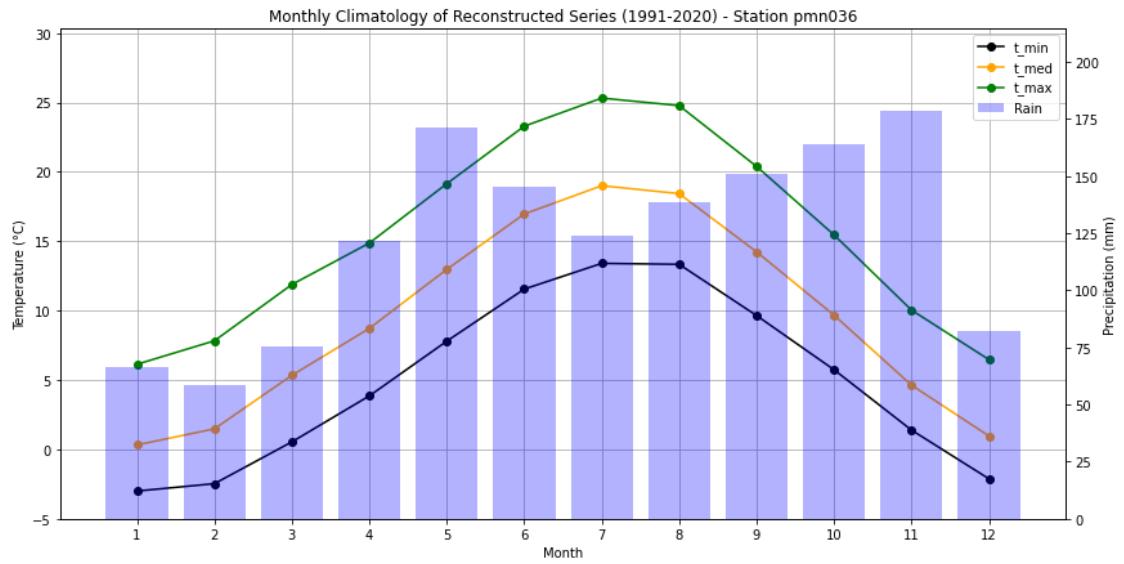


Figure A.30: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pmn036 during the period 1991-2020.

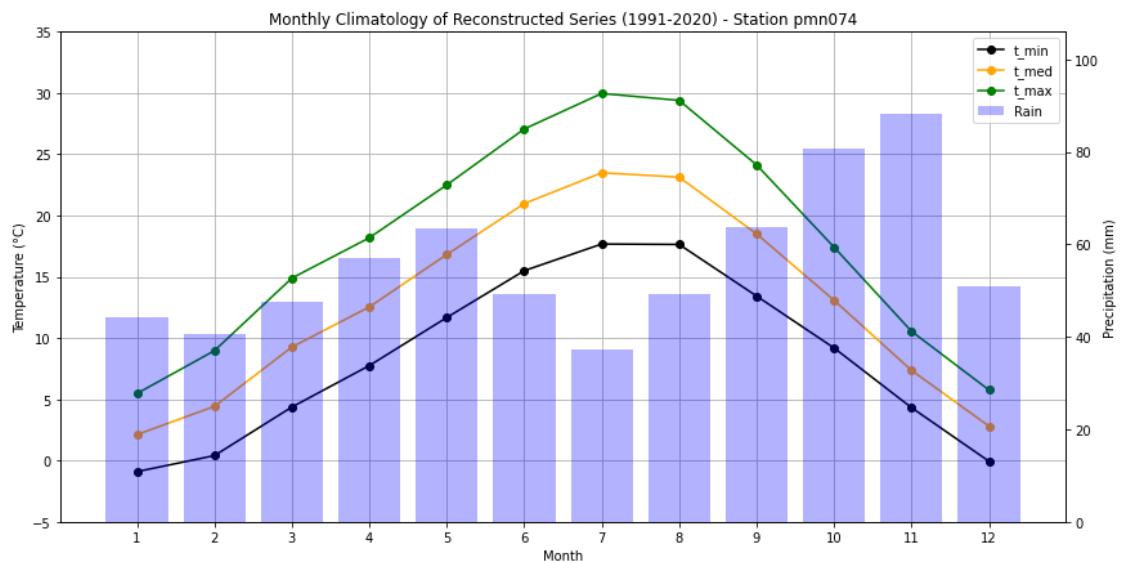


Figure A.31: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station pmn074 during the period 1991-2020.

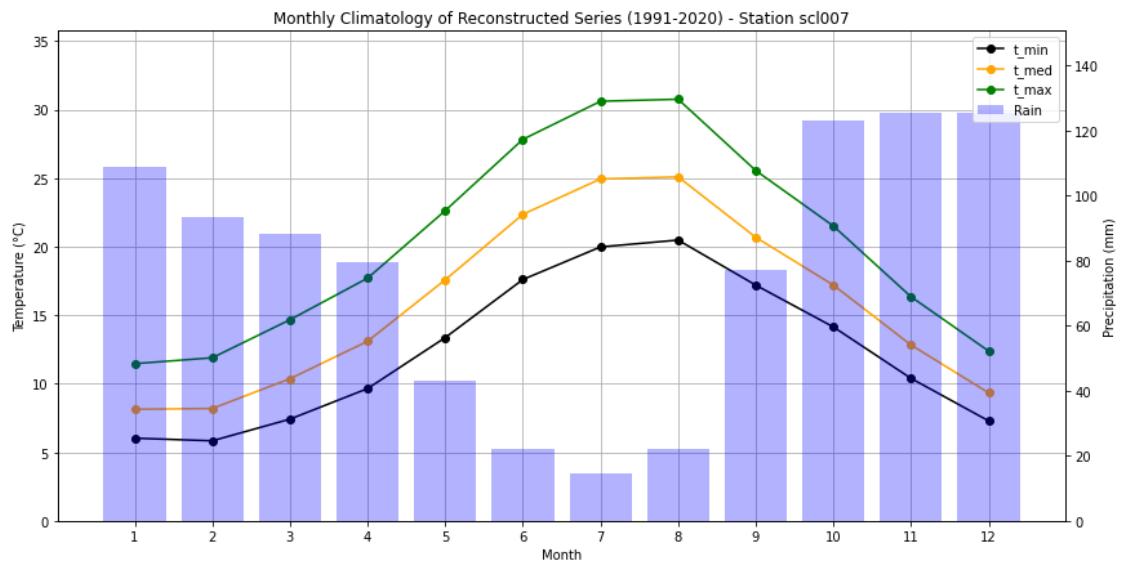


Figure A.32: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station scl007 during the period 1991-2020.

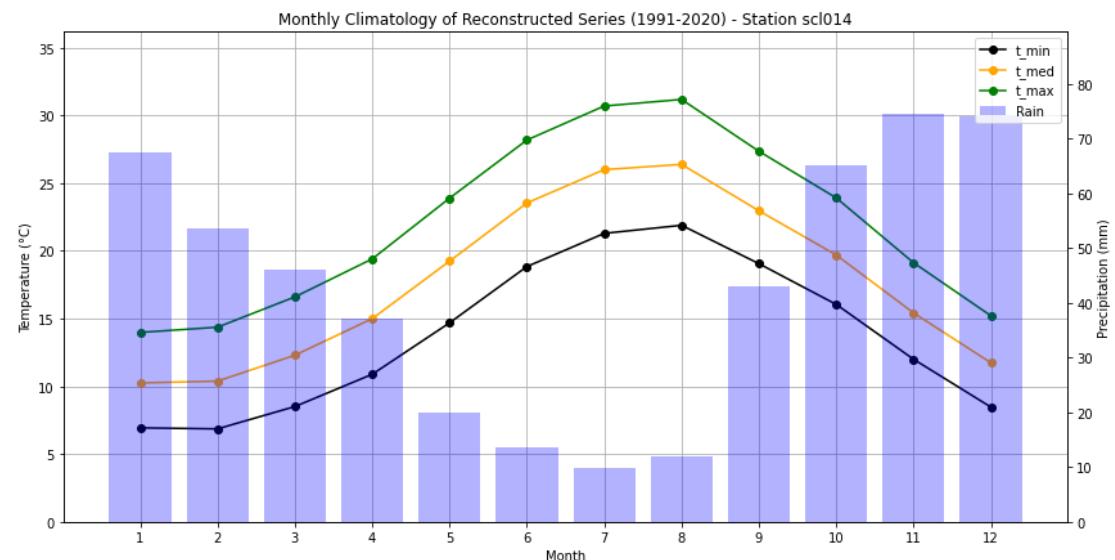


Figure A.33: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station scl014 during the period 1991-2020.

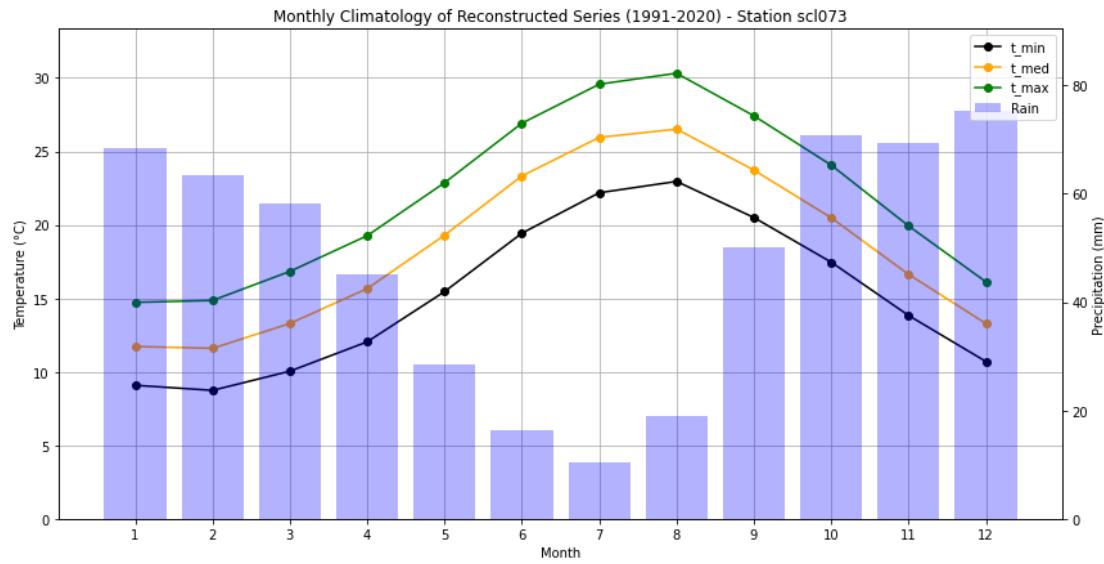


Figure A.34: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station scl073 during the period 1991-2020.

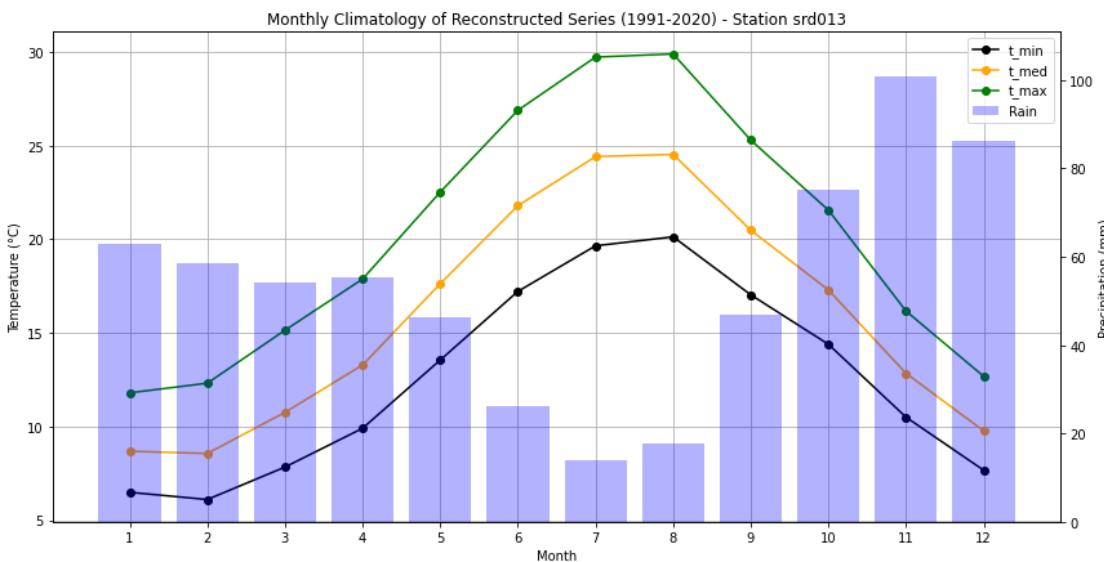


Figure A.35: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station srd013 during the period 1991-2020.

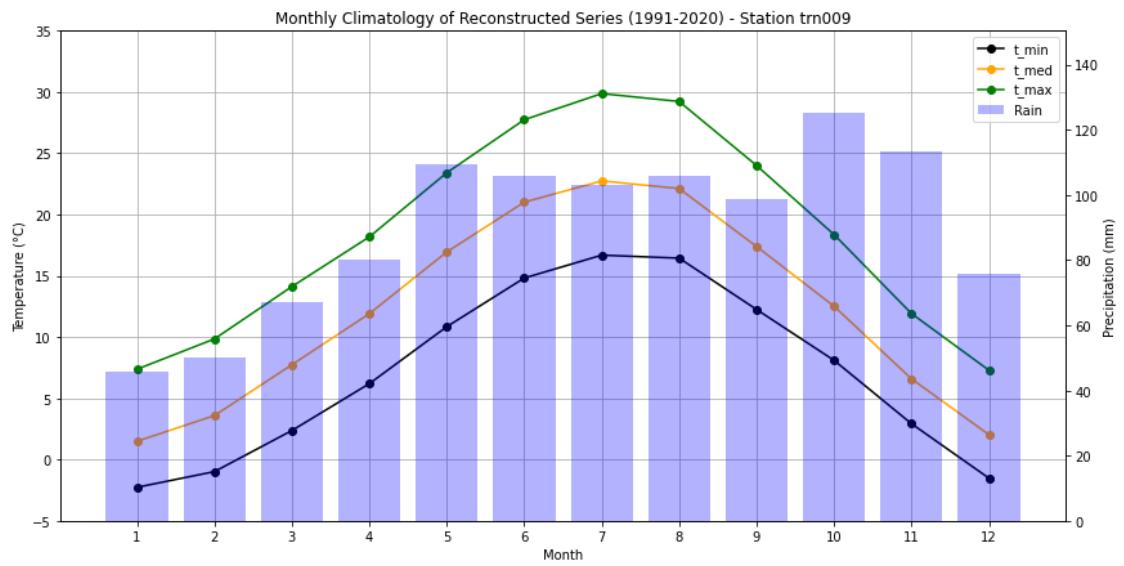


Figure A.36: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station trn009 during the period 1991-2020.

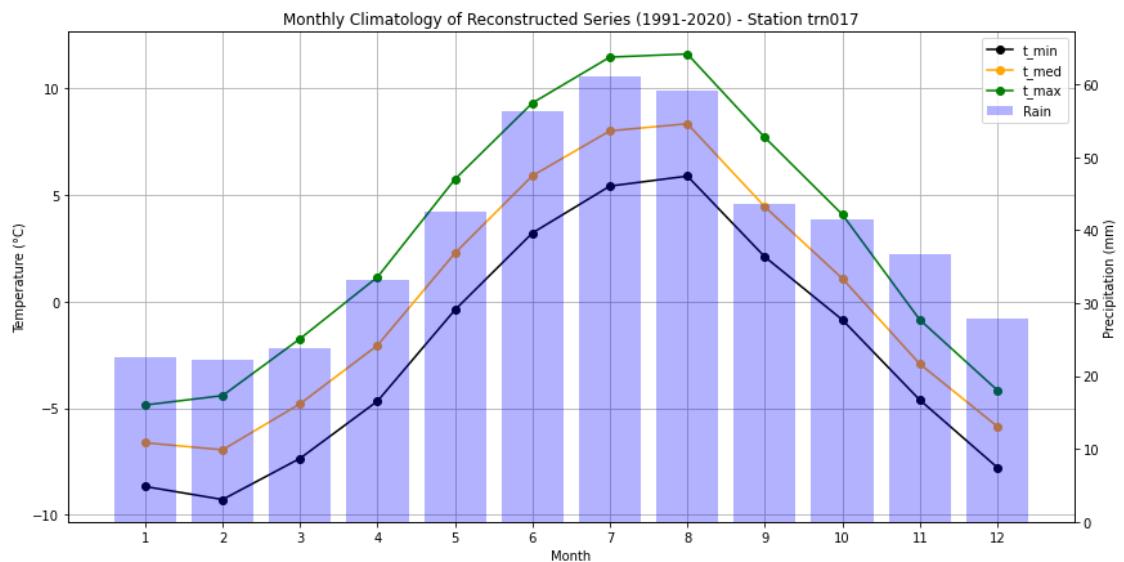


Figure A.37: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station trn017 during the period 1991-2020.

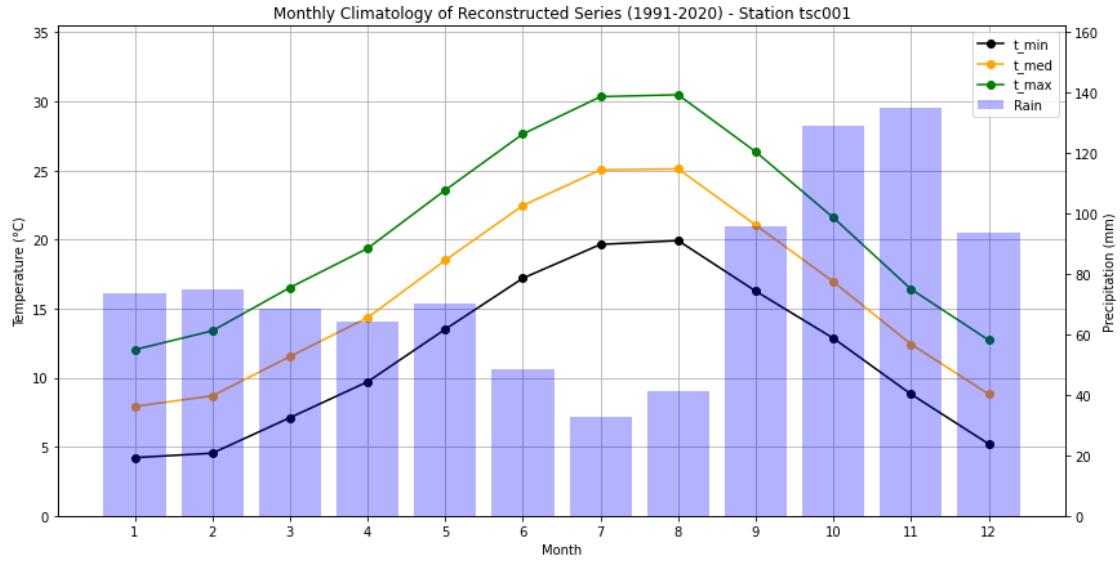


Figure A.38: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station tsc001 during the period 1991-2020.

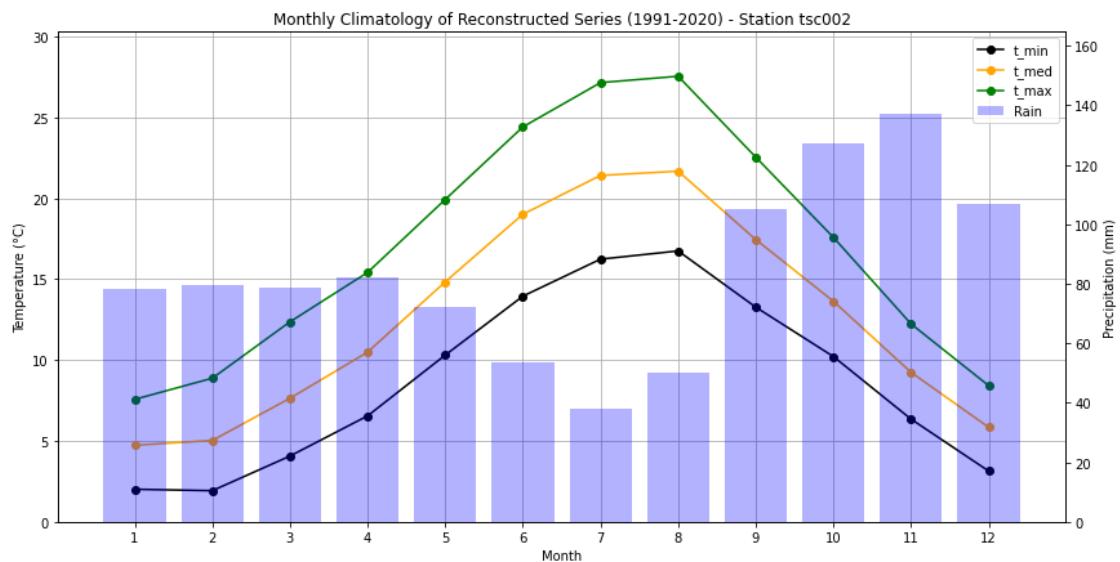


Figure A.39: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station tsc002 during the period 1991-2020.

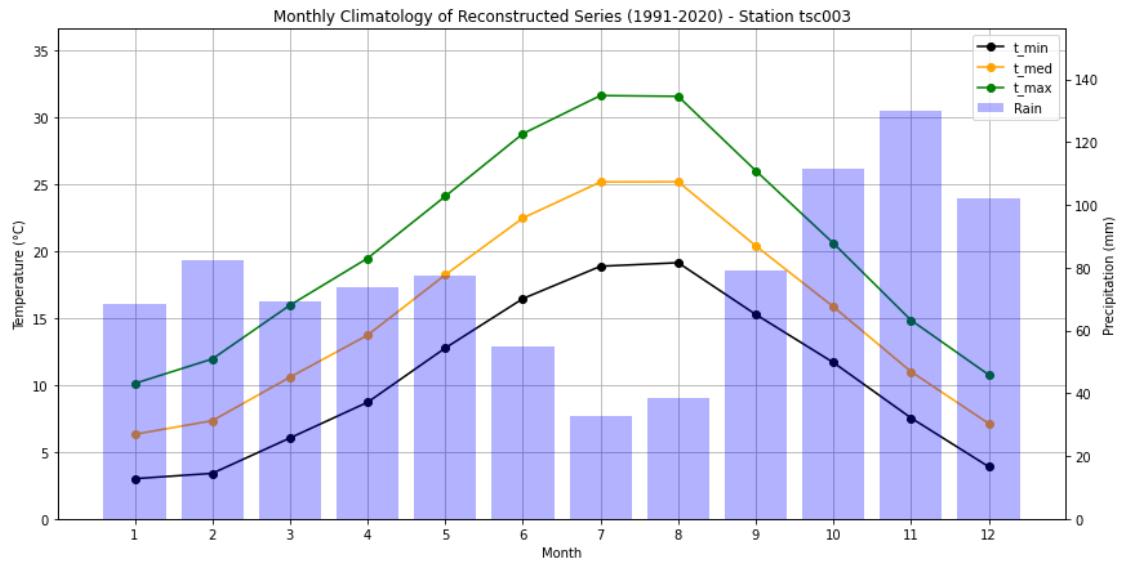


Figure A.40: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station tsc003 during the period 1991-2020.

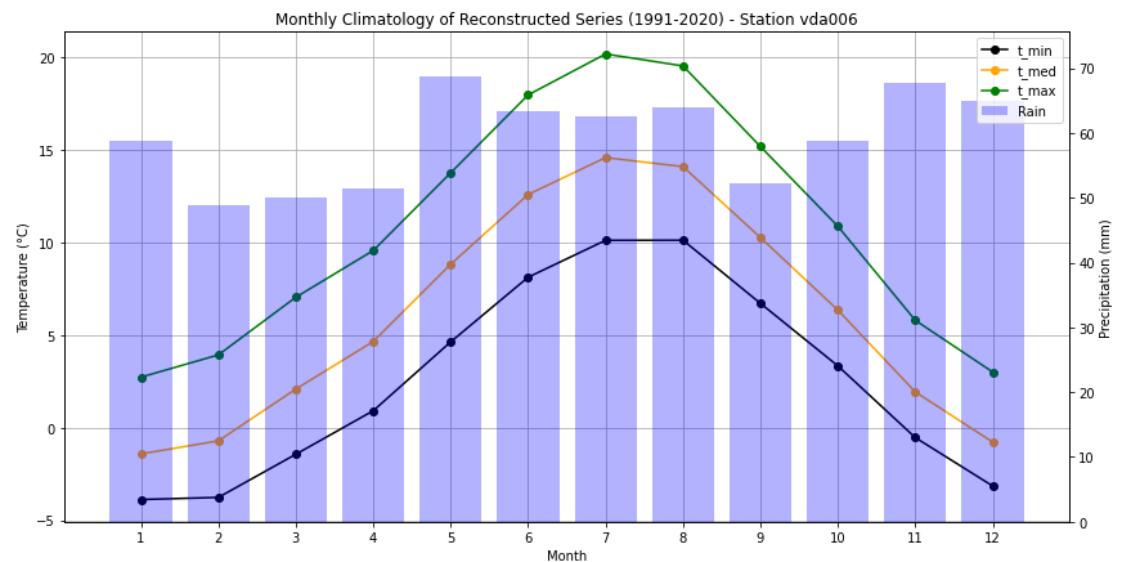


Figure A.41: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station vda006 during the period 1991-2020.

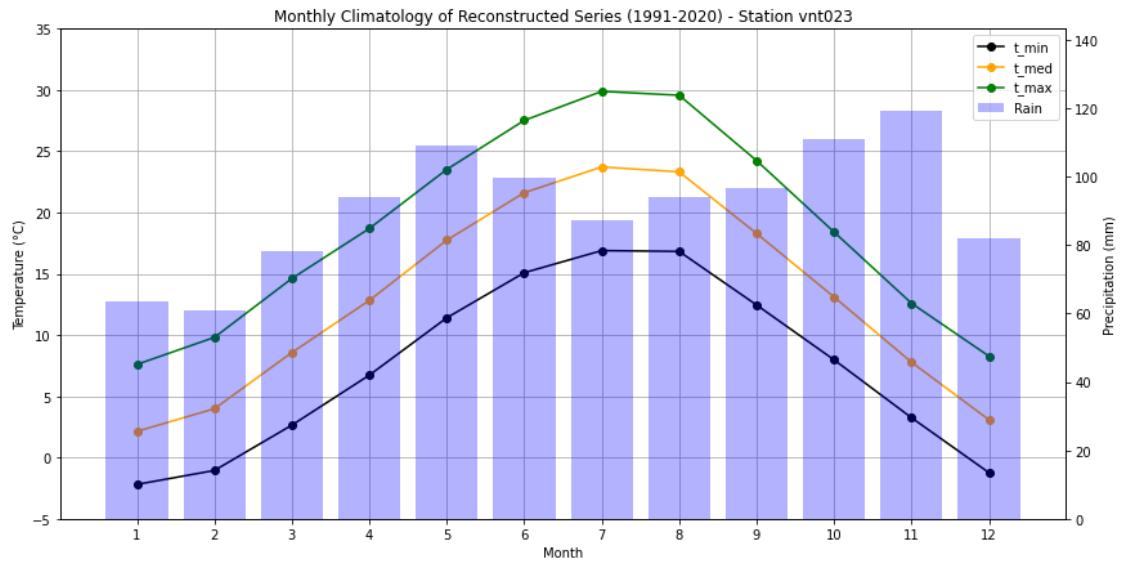


Figure A.42: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station vnt023 during the period 1991-2020.

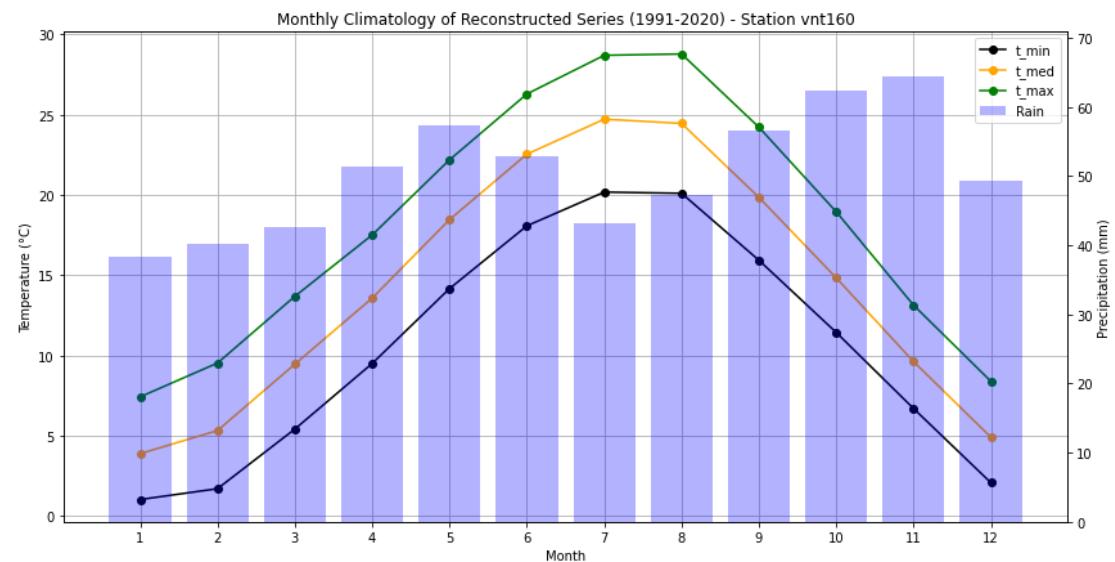


Figure A.43: Monthly climatologies of the reconstructed series of maximum, average and minimum temperature and precipitation of the station vnt160 during the period 1991-2020.

# List of Figures

Figure 2: Selected MNW observation station.....	16
Figure 5: Observation periods of maximum temperature of MNW stations.....	22
Figure 6: Observation periods of mean temperature of MNW stations.....	23
Figure 7: Observation periods of minimum temperature of MNW stations.....	23
Figure 8: Observation periods of precipitation of MNW stations.....	24
Figure 6: Mean biases of the average temperature of Era5-Land and CERRA.....	54
Figure 7: Mean biases of the average temperature of MERIDA H_RES and MSWX... ..	55
Figure 8: Average bias of the average temperature of VHR.....	55
Figure 9: Taylor diagrams considering only the nearest grid point.....	56
Figure 10: Taylor diagrams considering the nearest 16 grid points, aggregated with arithmetic mean.....	59
Figure 11: Taylor diagrams considering the nearest 16 grid points, aggregated with IDW.....	61
Figure 12: Taylor diagrams considering the nearest 16 grid points, with altitude correction and aggregated with IDW.....	62
Figure 13: Spatial distribution of the best resulting reanalysis dataset for precipitation and mean temperature.....	71
Figure 14: Spatial distribution of the best resulting reanalysis dataset for maximum and minimum temperature.....	71
Figure 15: Spatial distribution of the best resulting ML model for maximum, minimum and mean temperature.....	72
Figure 16: Monthly climatologies of the reconstructed series of station lmb021.....	80
Figure 17: Monthly climatologies of the reconstructed series of station lmb255.....	80
Figure 18: Comparison between monthly climatologies of temperature reconstructed using as input CERRA and ERA5-Land for the station lmb021.....	81
Figure 19: Comparison between monthly climatologies of temperature reconstructed using as input CERRA and ERA5-Land for the station lmb255.....	81

Figure 20: Comparison among monthly climatologies of temperature reconstructed using different types of ML models for station lmb021.....	82
Figure 21: Comparison among monthly climatologies of temperature reconstructed using different types of ML models for station lmb255.....	83
Figure 22: Comparison among monthly climatologies of precipitation reconstructed using different types of reanalysis datasets as input for station lmb021.....	83
Figure 23: Comparison among monthly climatologies of precipitation reconstructed different types of reanalysis datasets as input for station lmb255.....	84
Figure 24: Comparison between monthly reconstructed climatologies of temperature and the observed climatologies for station lmb021.....	85
Figure 25: Comparison between monthly reconstructed climatologies and the observed climatologies for station abr034.....	86
Figure 26: Comparison between monthly reconstructed climatologies of precipitation and the observed climatologies for station tsc001.....	87
Figure 27: Comparison between monthly climatologies of precipitation reconstructed and the observed climatologies for station ero001.....	87

## List of Tables

Table 4: Selected MNW observation stations and their characteristics.....	14
Table 5: Total number of observation days and days with missing data for each station for the variable mean temperature.....	18
Table 6: Total number of observation days and days with missing data for each station for the variable precipitation.....	20
Table 4: Taylor diagram statistics for the CERRA reanalysis of the variable T avg, considering 1 and 16 points.....	58
Table 5: Taylor diagram statistics for the MSWEP reanalysis of the variable P, considering 1 and 16 points.....	58
Table 6: Taylor diagram statistics for the CERRA reanalysis of the variable T avg, considering 16 points aggregated with arithmetic mean and IDW.....	60
Table 7: Taylor diagram statistics for the MSWEP reanalysis of the variable P, considering 16 points aggregated with arithmetic mean and IDW.....	60
Table 8: Taylor diagram statistics for the CERRA reanalysis of the variable average temperature, considering 16 points with and without elevation correction and aggregated with IDW.....	62
Table 9: Performance of different type of ML models for mean temperature of station lmb021 using as input CERRA.....	66
Table 10: Performance of different type of ML models of variable T max of station lmb021 using as input CERRA.....	66
Table 11: Performance of different type of ML models of variable T min of station lmb021 using as input CERRA.....	66
Table 12: Performance of different reanalysis of variable T avg of station lmb021 using the model SVR with linear kernel.....	66
Table 13: Performance of different reanalysis of variable T max of station lmb021 using the model SVR with linear kernel.....	67
Table 14: Performance of different reanalysis of variable T min of station lmb021 using the model SVR with linear kernel.....	67
Table 15: Performance of different reanalysis of variable P of station lmb021 using the model SVR with radial kernel.....	67

Table 16: Best reanalysis model for each station and variable.....	68
Table 17: Best ML model for each station and variable.....	69
Table 18: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable T avg for each station.....	73
Table 19: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable T min for each station.....	74
Table 20: Performance of the SVR models with linear kernel and CERRA as input, used to reconstruct the synthetic series of the variable T max for each station.....	75
Table 21: Performance of the SVR models with radial kernel and MSWEP as input, used to reconstruct the synthetic series of the variable P for each station.....	77

## Acknowledgments

Desidero innanzitutto ringraziare il professor Ceppi e Guido per la loro grande disponibilità e per avermi guidato con pazienza e professionalità durante la stesura di questo lavoro.

Un ringraziamento speciale va a mio papà, che è sempre stato la mia costante e il mio punto di riferimento. Grazie per avermi sostenuto in ogni momento ed essermi sempre stato accanto.

Un pensiero va anche a mia mamma, che, sebbene non sia più qui per condividere questo traguardo, so che mi è sempre stata vicina. Mi manchi ogni giorno, e spero che tu sia fiera di me.

Un particolare grazie va a Giovanna, che in questi ultimi dieci anni mi ha sempre incoraggiato e sostenuto. Mi hai aiutato a crescere e a credere in me stessa, anche quando era difficile farlo.

Ringrazio di cuore i miei nonni, che mi hanno sempre dimostrato un amore infinito e incondizionato.

Un ringraziamento speciale va a Martina e Noemi. Mi siete state vicine sia nei momenti più difficili che in quelli più belli. La pallavolo mi ha dato la fortuna di incontrare persone meravigliose come voi, con cui posso essere sempre me stessa, sapendo di essere accettata.

Un grazie speciale va a Elisa e Sara. Senza compagne come voi, questo percorso sarebbe stato molto più difficile, non sarei arrivata fin qui senza il vostro costante supporto.

Vorrei ringraziare anche Francesca, Simona e Fabiola, per avermi sempre regalato un sorriso, anche nei momenti più bui.

Infine, vorrei ringraziare le famiglie Bagnati e Bovio al completo. Mi conoscete da quando sono nata e mi avete sempre accolto come parte delle vostre famiglie. Per me siete come una seconda famiglia, e vi sarò sempre grata per questo.

