Università degli Studi di Torino

Dipartimento di Fisica

Corso di Laurea Margistrale in Fisica

Data-driven definition of the meteorological seasons and their expected changes in the 21st century

Tesi di Laurea Magistrale

Supervisor:
Prof.ssa Elisa Palazzi

Candidate:
Jacopo Grassi

Co-supervisor:
Dott. Paolo Davini

Anno accademico 2021/2022

The work presented in this thesis was performed at the Institute of Atmospheric Sciences and Climate (ISAC) of the Italian National Research Council (CNR), Torino (corso Fiume, 4). The whole research has been carefully supervised by Dr. Paolo Davini, researcher at CNR-ISAC and by Prof. Elisa Palazzi, University of Turin.

Contents

Chapter 1

Introduction: what are the seasons?

The concept of *seasons* is something that everyone has in mind, since it is part of daily life. On the other hand, giving a shared and unique definition of what the seasons are seems a harder task. Firstly, because there is not only a type of seasons: there are astronomical seasons, meteorological seasons, but also the flu season, the high season of a tourist destination, and so on. We are used to associate the concept of seasons to everything that shows a certain periodicity, which we call *seasonality*. But, if this periodicity is the seasonality, what is the formal definition of seasons? This lack of clarity seems not to be confined to terminology and leads to an ambiguity that can become limiting when, for example, we wonder what will happen to the seasons in the future. A first approach for trying to give a shape to the concept of seasons should start from seasonality itself.

Seasonality is a wide concept which affects many aspects of everyday life. However, the definition of seasonality is not straightforward: it could be defined in a general way as a recurring pattern or even cycle that occurs at regular intervals within a specific time frame, and it could be observed in a wide range of natural and human made phenomena. Earth science disciplines (meteorology, botany, glaciology, etc. etc.), but also economy, finance, epidemiology, and a wide range of other sectors, show seasonal patterns.

The correct identification of seasonal patterns is a crucial step when dealing with phenomena related to several of the above-mentioned sectors. A correct identification of seasonality allows us to deeply investigate the features of the phenomenon we are studying, and to achieve a better understanding of it. This has multiple positive repercussions, since it allows us to make better decisions and develop adequate strategies for the specific problem, or even develop methods for forecasting.

While seasonality is a property which shows itself in the manifestation of a

phenomenon, its source is often more difficult to investigate, as it is the result of many mutually interacting factors. In meteorology, seasonality is the tangible demonstration of the earth axis obliquity with respect to the rotation plane. In the economy, e.g. in tourism, seasonal patterns are affected by climate and weather, social customs (e. g. holiday periods), business customs and need for supply. In epidemiology seasonality leads to the propagation of flu and other pathogens which is determined by biological, social, and environmental factors. Thus, a first approach to seasonality problems is often done investigating the behavior of the phenomenon, rather than its causes. This approach is the so-called time series analysis.

In meteorological and climatic sciences seasonality is an element which plays a central role. Being an essential element of Earth's climate system, seasonality is used to characterize the climate of different regions. Also projections of future evolution of climate are analyzed looking at changes of the seasonal pattern. The ability to reproduce seasonal patterns is also used for the validation of climate models, i.e., numerical models which simulate the behavior of Earth's climate system.

Especially when dealing with meteorology and climate, we often refer to seasonality using the concept of seasons. Although this could seem only a matter of terminology, it is not. The seasons are the periods in which we artificially divide the year, and last typically three months. Seasonality, being an oscillation, is typically modeled using continuous functions, such as sinusoidal functions. This means that somewhere stands the assumption that dividing the year in seasons is a good way to describe the Earth's climate system seasonality.

It is worth pointing out that in everyday language, when speaking of weather and climate, we often use the word *seasons* ambiguously, referring to what are technically called *astronomical seasons*. The astronomical seasons are defined on rigorous criteria, based on geometrical factors of rotation and revolution of the Earth. The different behaviour of the meteorological weather in different periods of the year is described by the so-called *meteorological seasons*. Astronomical and meteorological seasons are strictly related: the inclination of the rotation axis with respect to the rotation plan determines in different periods of the year a different distribution of the solar radiation through the Earth's surface. Nevertheless, meteorological and astronomical seasons could not be treated as a single entity, since the response of the weather at different insolations can vary considerably depending on the locality. In this work we will use the word *seasons* referring to the meteorological seasons.

After this necessary digression on the nomenclature, we can go back to wondering about the link between seasonality and meteorological seasons. Looking at it from another point of view, if each of the 365 days in a year typically has its own climatic behavior due to seasonality in a specific locality, then seasonality could be described

by 365 values. The definition of the seasons implicitly assumes that the information contained in these 365 values could be condensed in (typically) four values, one for each season. Intuitively, this could seem a good description of the seasonal cycle. If we look, for example, at the behavior of atmospheric conditions in January 1^{st} of any one year, and compare it to the behavior of January 1^{st} of another year, we will probably observe two different conditions. This is due to the variability of the system at high frequencies. On the other hand, we can take all the January $1^{st}s$ in a sufficiently wide range of years (typically 30 years, and in this case we are talking about Climatology) and obtain a distribution of possible conditions for the first day of the year. We will probably be able to observe that this distribution is quite similar to the one obtained taking all the January $15^{th}s$, or January $30^{th}s$, but different to the one obtained taking July 1^{st}. Thus, grouping days with similar distributions seems a logical approach.

Although, despite the extensive use we make of them, meteorological seasons are more an heuristic concept than well defined entities. This is primarily due to the fact that it is not possible to give a globally valid definition of meteorological seasons. In fact seasonal patterns vary according to the locality (e.g. orography, vegetation, prevailing winds, etc.) and the specific physical variables that we take into account. At midlatitudes, we often consider four seasons looking at temperatures: a hot one (Summer), a cold one (Winter) and two transition seasons (Spring and Autumn). In subtropical areas subjected to monsoonal dynamics, the seasonal division is performed distinguishing between the monsoonal wet season and the dry season And many other example could be found. These seasonal divisions are most of the times based on heuristic consideration. When they are performed on more rigorous criteria, it requires a long work for the identification of the physical variables and threshold to be taken into account. Furthermore, the division into seasons usually rarely recognizes a time resolution inferior to one month. The last problem we point out is that the same division into seasons that is used nowadays, is always used when analyzing the future projection obtained by climate models. There is multiple evidence that in the last decades a wide range of seasonal patterns has changed, and we can then assume that also the division into seasons should be constantly verified and updated.

The first purpose of this work is to develop a methodology for the identification of meteorological seasons in climatic datasets, trying to minimize the arbitrary assumptions. This methodology will be constructed with the aim of being as general as possible and consequently applicable to the most disparate cases. As we highlighted before, approaching the problem of the division in seasons in a physically-driven way shows many difficulties related to the variability with which seasonal patterns show

in different areas. Thus, we will try to use a data-driven approach, making use of a series of machine learning tools. As we will explain, the power of machine-learning relies on the fact that a set of algorithms could be instructed to autonomously recognise the best criteria to use for the division in seasons.

Our second purpose is to find a way for evaluating how the seasons detected are represented in different climate datasets. This would result in at least two insightful applications. One the one hand, this will provide us of a tool for evaluating how different datasets represent the seasons. For example, this can be used to robustly analyze the presence of bias in the representation of seasonal cycle in climate models . In light of what we said before, this could be a powerful tool in the validation of climate models. On the other hand, this will allow us to study to what extent the seasons we are experimenting nowadays are expected to change in the future, making use of the projection made by climate models.

Both of our goals are pursued by trying to develop a general methodology. Such methodology will be applied to a selected case study making use of total precipitation and the surface air temperature, which are the two most used variables for climatic characterizations, to define seasons.

This thesis is structured as follows. Firstly, in Chapter 2, we give a formal overview of time series analysis and of the required assumptions in order to defining the seasons starting from the concept of seasonality. Then, in Chapter 3, we describe the design of the machine learning algorithms that we chose for achieving our purposes. Chapter 4 provides a brief description of the datasets we will use for a first application of our methodology. These datasets are the ECMWF ERA5 Renalysis for the recognition of seasons in the present and recent past, and the EC-Earth3 Earth System Model for the tracking of seasons evolution. Finally, in Chapter 5 we apply our methodology to a case study, which is the Hindu Kush Karakoram Himalaya region, in the northern part of the Indian subcontinent. The main scientific questions we will try to address in this part are:

- Is the division in seasons that we are currently using correct? We will try to answer this question both about the number of seasons and the dates we use for the splitting.

- How are the seasons recognised by machine learning represented in a climate model?

- How are these seasons expected to change in the future?

It is worth noting that this work represents an original attempt to apply methods taken from different domains to a problem that is not well documented in the

literature, such as the definition of the meteorological seasons. Thus, we will try to present the results obtained in this dissertation with a special focus on what can be improved or needs further investigation.

Chapter 2

Seasons and seasonality

As seen in the previous chapter, there is a lack of shared and uniform definitions of meteorological seasons, and it is not easy to find a common ground even to define methods for their recognition. This is mainly due to the differences among seasons behavior in different regions of the world, in timing, amplitude of the signal and involved variables. In this work we will not try to fill this lack of definitions in an exhaustive way, being generality and flexibility the main objectives we want to achieve. Nevertheless, some general criteria and assumptions about what makes seasons distinguishable elements of the climate system must be introduced, in order to have a starting point for the construction of our methodology. In this chapter we will try to formulate a work hypothesis, which relies on some basic concepts of time series analysis. Section 2.2 contains a formal treatment about seasonality in time series analysis, seen as a deterministic signal. In section 2.3 the application to multidimensional climatic data is presented, along with the assumptions which allow us to transform this continuous signal into a finite number of similar periods within them, which would be the seasons.

2.1 Seasonality on time series analysis

Since seasonality is a wide concept which affects phenomena in a wide range of sectors, there are many ways to approach its evaluation. Neglecting the source of seasonality and focusing on its phenomenology is the approach used in the so-called time series analysis. Nevertheless, seasonality is not the only component present in time series, and it is not possible to focus on seasonality completely neglecting the other components. Thus, the time series components which are taken into account and the way in which they are investigated could vary depending on the purpose of the analysis.

In this section we will try to give a theoretical overview of the basic principles of time series analysis, favoring the point of view that is usually held for the analysis of climatic time series. The main focus in this part is the recognition of seasonality. Thus, we will prefer an approach focused on seasonality rather than on formal completeness. This part is mainly inspired by [Hamilton, 1994] and [NIST, 2012], which could be used as references for a complete treatment.

2.1.1 Basic definitions

Consider a time-series X of T real values generated by a stochastic process A:

$$X = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}, \quad x_t \in R^K \tag{2.1}$$

Assume each x_t to be a particular realization of a generic probability density function f_t, which is determined by an undefined set of parameters γ_t: we will denote the probability of getting x_t by f_t as $f_t(x_t|\gamma_t)$.

Consider now an ensemble E of N time-series generated by the same process A:

$$E = \{X^1, X^2, ..., X^N, ..., X^N\} \tag{2.2}$$

It's convenient to rearrange data in a TXN matrix D where each row represents a time step and each column a time series:

$$D = \begin{bmatrix} x_1^1 & \ldots & x_1^N \\ \vdots & \ddots & \vdots \\ x_T^1 & \ldots & x_T^N \end{bmatrix} \tag{2.3}$$

We can easily observe that if $K > 1$ then D is a three-dimensional tensor. Each row of D is thus ruled by the same probability density function f_t, and we can summarize these functions in a column vector F:

$$F = \begin{bmatrix} f_1(x_1|\gamma_1) \\ \vdots \\ f_T(x_T|\gamma_T) \end{bmatrix} \tag{2.4}$$

The simultaneous knowledge of all the functions f_t and the parameters γ_t gives the statistic of the time series generated by the process A, which is the goal of a parametric approach. This allows to compute some useful indicators about the time series, which could be arranged in column vectors:

- **Expectation** (or ensemble mean):

$$\mu_t = E[X_t] = \int_{-\infty}^{\infty} x_t f_t(x_t|\gamma_t) dx_t \tag{2.5}$$

Each row of F has its own expectation value, thus we can define the column vector M of these values, that will represent the ensemble mean at each timestep:

$$M = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_T \end{bmatrix} \tag{2.6}$$

- **Central moments**:

$$\sigma_t^k = E[(X_t - \mu_t)^k] = \int_{-\infty}^{\infty} (x_t - \mu_t)^k f_t(x_t|\gamma_t) dx_t \tag{2.7}$$

As for M, thus we can define the column vector S^k of the moment of each row of F:

$$S^k = \begin{bmatrix} \sigma_1^k \\ \vdots \\ \sigma_T^k \end{bmatrix} \tag{2.8}$$

- **Autocovariance sequence**:

$$\rho_{t,t-l} = E[(X_t - \mu_t)(X_{t-l} - \mu_{t-l})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_t - \mu_t)(x_{t-l} - \mu_{t-l}) f_t(x_t|\gamma_t) dx_t \tag{2.9}$$

The autocovariance is defined for each row and for each possible lag, and thus the result could be summarized in a column vector R of length $T * T$:

$$R = \begin{bmatrix} \rho_{1,1} \\ \rho_{1,2} \\ \vdots \\ \rho_{1,T} \\ \rho_{2,1} \\ \vdots \\ \rho_{T,T} \end{bmatrix} \tag{2.10}$$

If we knew the explicit form of F, we would be able to perform a complete analysis of our time series. In fact, the information contained in F would allow us to know the statistical behavior of the process A at each time step. In this scenario, also seasonality would be described by the statistic.

The first problem is that when approaching the analysis of a time-series, the *a priori* knowledge of both f_t and γ_t is most of the time unsatisfied. A statistical approach could be supposing the shape of functions f_t and then estimating the best parameters γ_t. This would require the ensemble of realizations E. Here the second problem raise: dealing with climatic data, especially when dealing with observations, we usually only have one realization X and not the ensemble E.

These problems could be bypassed with the concepts of stationarity and ergodicity, which will be detailed in the next section. Briefly, we can assume that our time series shows certain stability characteristics which allow us to treat a single realization as an ensemble realization.

2.1.2 Stationarity and ergodicity

As said before, when the ensemble of realization of a process A is not available, the way we have to statistically investigate the properties of a time series is suppose that the time series itself could be treated as an ensemble realization. This is made

using the concept of stationarity, in different grades. As we will see with the formal requirements, stationarity is an assumption. In fact, we assume that if we had an ensemble realization, it would behave in a certain way.

A process is said to be strictly stationary if each time step shows the same statistical behavior. Formally:

$$f_{t1}(x_{t1}|\gamma_{t1}) = f_{t2}(x_{t2}|\gamma_{t2}) \quad \forall x_{t1}, x_{t2} \tag{2.11}$$

Which is equivalent to require F to be constant, and consequently M, S^k and R. This definition is too strong for most of the interesting time-series, and not applicable to climate data. Furthermore, a time series which has, by definition, no statistical difference between different timesteps, obviously loses the interesting features we are looking for, such as seasonality.

There is a wider condition we can impose to time series, called Wide-Sense Stationarity (WSS). A process is said to be WSS if:

1. For each time step the expectation value is constant:

$$\mu_{t1} = \mu_{t2} := \mu \quad \forall t1, t2 \tag{2.12}$$

2. The expectation value of the squared signal is finite:

$$E[|x_t|^2] < \infty \quad \forall t \tag{2.13}$$

3. The autocovariance sequence varies only in function of the lag:

$$\rho_{xx}(t1, t2) = \rho_{xx}(t1 - t2, 0) \quad \forall t1, t2 \tag{2.14}$$

A WSS process shows a certain regularity between the statistical behaviour of different timesteps. This regularity is wider than the one imposed by strictly stationarity since, for example, central moments could vary through different timesteps.

WSS allows to collapse the statistical indicator of the time series. M could be collapsed in a single value, and R, being a function only of the lag, could be reduced in a $2 * T - 1$ column vector. Since R is symmetric respect to $l = 0$, it is possible to consider only one side and redefine:

$$R = \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_T \end{bmatrix} \tag{2.15}$$

The main advantage of strict stationarity and WSS relies in the fact that we can consistently redefine the statistical indicators shown before (expectation, central moments, and autocovariance) using the average on time dimension. Considering our starting time series X we can define:

$$\overline{x} = \frac{1}{T} \sum_{t=0}^{T} x_t \tag{2.16}$$

$$\overline{\sigma^k} = \frac{1}{T} \sum_{t=0}^{T} (x_t - \overline{x})^k \tag{2.17}$$

$$\overline{\rho^k}(l) = \frac{1}{T-l} \sum_{t=0}^{T-l} (x_t - \overline{x})(x_{t-l} - \overline{x}) \tag{2.18}$$

Where the overbar indicates that the average is computed on time dimension and not on ensemble realizations. A further step is ergodicity. Even if the operation 2.16, 2.19, and 2.18 are consistent, we cannot state that they are equivalent to expectation, central moments, and autocovariance defined in 2.5, 2.7, and 2.9 on the ensemble realizations. When it happens, the process is said to be ergodic. Such as for stationarity, there are different grades of ergodicity. The most used in time series analysis is ergodicity for the mean.

Formally, the process A is said to be ergodic for the mean if:

$$\lim_{T \to +\infty} \overline{x} = \mu \tag{2.19}$$

i. e., if the mean on time dimension converges to the ensemble mean.

It worth note that in many applications ergodicity and stationarity turn out to amount in the same requests, but they are different concepts, as detailed before. While ergodicity is a sufficient condition for stationarity, a stationary process could not be ergodic. Furthermore, it must be clear that stationarity and ergodicity are most of the time assumptions, being used when the ensemble realization is not available.

We wonder now if stationarity and ergodicity are concepts compatible with the presence of seasonality. There is no a general answer to this question, since it depends on how seasonality influence the statistical behavior of each timestep. Strict stationarity is too strong for allowing seasonality, but we already stated that is too strong for any interesting application. WSS could be compatible with seasonality if, for example, seasonal patterns only influences central moments of each timestep's statistical distribution. This requirements is too strict, and we must take into account that seasonality in climate system could affect the expectation values too, contrasting the requirement in equation 2.12.

Thus, for the evaluation of seasonality in Earth's climate system we cannot rely on statistical methods. The most used alternative is to rely on some heuristic assumptions, dividing the time series in components and trying to detect them with *ad-hoc* defined methods. These considerations will be discussed in the following subsections.

Even if stationarity and ergodicy did not lead to usable results, this dissertation will be useful in the following of this chapter.

2.1.3 Sampling

The data considered since now are by definition discrete in time. Even if the process that generates them is time-continuous, a discretization process must be applied to access the data. This process is called sampling and is performed using a sampling interval Δt. So given a continuous signal $x(t)$ the sampling process could be formalized:

$$x_t \to x_t = x(n\Delta t) \tag{2.20}$$

Where $n \in [1, T]$ is the number of measurements. The sampling process determines a loss of information and must be chosen carefully, according to the phe-

nomenon which is under investigation. The choice of a sampling interval implies the definition of a sampling frequency ν, defined as:

$$\nu = \frac{1}{\Delta t} \tag{2.21}$$

Here it must be noticed that the definition of the unit of measure of t is not trivial, since it implies the assumption of a reference time unit which is a system scale factor.

In this work we will assume $t \in]0, 1[$ such that $\nu \in I^+$ where I^+ denotes the integer numbers larger than one. This means that ν describes the number of samplings taken in a time unit. For simplicity we introduce now the parameters used in this work: the reference time unit is 1 yr with a daily frequency sampling ($\Delta t = 1/365$), thus ν=365. This means that we will ignore processes that are characterized by frequency higher than 2 days.

2.1.4 Time series components

As stated in previous subsections, in absence of the ensemble realization of the process, seasonality in time series could not be evaluated relying on statistical methods. Thus, we can try to divide the time series in its components and focus on seasonality.

Time series could be heuristically considered as formed by three components: trends (T), cycles (C), and residuals (R). These components are usually modeled combining them in additive or multiplicative ways. Here we consider the addictive mode:

$$X = T + C + R \tag{2.22}$$

A finest decomposition could be performed on cyclical components. Cycles could include an oscillation with period inferior or equal to 1 yr (seasonality Se) and lower frequencies components (Cy). It is not merely a matter of periodicity. Seasonal cycles in climate system are usually more regular than other cyclical components. In the same way, climatic time series usually shows variability at short time scales, such as days or weeks, which could lead to rise of irregular periodicity too.

Determining threshold of regularity or periodicity for the division of periodical or quasi-periodical cycles in seasonality or other components would requires additional evaluations which exceed the purpose of this work.

Here we are trying to find the point of contact between the definitions of seasonality and seasons. Meteorological seasons, in their common use, have a periodicity of about 1 yr. Thus we will consider the cycles with period of about 1 yr as seasonal components (Se), and we will incorporate the other periodicity or quasi-periodicity into residuals (R). This results in the following division of time series into its components:

$$X = T + Se + R \tag{2.23}$$

Time series analysis makes use of different techniques in order to identify these components, based on the definition of continuous functions which can represent them. Assuming that we could chose the best one, we would be able to find the seasonal pattern in our climatic time series. This pattern would be described by a time-contiguous signal. Thus, this does not answer our main question, which is why we can use meteorological seasons for the description of the seasonal pattern in climate time series. At the light of what we stated in this section, we will try to answer this question in the next section.

2.2 Definition of seasons

Briefly summarizing the results of the previous section: a pure statistical approach for the recognition of seasonality is not applicable. Firstly, because most of the time, when dealing with climate time series, we only have one realization and not an ensemble. Furthermore, the intensity of seasonal patterns in climate time series prevents us from assuming stationarity, in order to use statistical approaches with only one realization. Thus, we stated that the best approach is to model seasonality assuming an heuristic division of time series in components. In this sense, seasonality should be modeled using a time-continuous signal. This is not helpful for the definition of seasons, meant as the periods in which we divide the years for describing seasonality.

As we stated in chapter 1, meteorological seasons are a powerful tool since they describe what is the expected behavior in a determined period of the year. Thus, we will try to combine the heuristic division performed in , with the ensemble approach. As we will detail, using eqaution 2.23 we can obtain an ensemble of realizations starting from a single realization, which statistical behavior highlights the seasonal features.

2.2.1 Work approach

Let's consider a time series which is a single output of a stochastic process. We do not have an ensemble of realizations, and we can not assume stationarity. Nevertheless, we can consider each year as a single output of the process. If the time series, being daily the frequency sampling, consists of $T\,years * 365$ values, now we have a matrix of $T\,years$ realizations of the process A, each of one of length 365. This matrix is formally consistent with D (equation 2.3), but substantially different if we consider that the components defined in section 2.1.4 now vary through each realization. In fact, each row now has a time-dependent statistic. That is to say, the first realization (i.e., the first year) is not necessary ruled by the same distribution on the n^{th} realization (i.e., the n^{th} year). Basing on the simplified time series components division performed in equation 2.23 we can note that the trend could modify these distributions. If we remove the trend, as we will see in chapter Methods, we can consider our representation formally more reliable.

It is easiest to get this point considering an ideal climatic dataset. Climatic data could be presented in spatio-temporal matrices. Consider a homogeneous space-time distributed dataset, where each space coordinate indicates a grid point, and a defined number of atmospheric variables. Assuming that there are M grid points and H variables, the size of the matrix is:

$$(T\,years * 365)\,X\,(M * H) \tag{2.24}$$

If we consider each year as a different realization of the process, we obtain a tensor with the following size:

$$(365)\,X\,(T\,year)\,X\,(M * H) \tag{2.25}$$

We can represent the data in the matrix D_f, remembering that each $x \in R^{M*H}$:

$$D_f = \begin{bmatrix} x_1^1 & \dots & x_1^T \\ \vdots & \ddots & \vdots \\ x_{365}^1 & \dots & x_{365}^T \end{bmatrix} \tag{2.26}$$

It should be now clear that the current data representation (after having removed eventual trends) allows us to focus only on seasonality and residuals components.

With this representation, each rows contains the climatic behavior for each day on different years, grid points and variables.

This representation is now compatible with our purpose of defining the seasons. As we will detail in the next subsection, finding meteorological seasons is equivalent to grouping the rows in D_f basing on their distributions.

2.2.2 How to define the seasons

Now we can formalize our work hypothesis for the division in meteorological seasons. At the light of what we said in the previous sections, we can assume that seasonality in climate system components shows an emergent behavior which leads to the identification of periods with similar characteristics, i. e. the seasons. We will verify the goodness of this hypothesis at the end of this dissertation.

Formally, this could be seen as a mapping from a time continuous signal to a discrete number of states. Consider the matrix D_f defined in the previous section (equation 2.26). Consider taking just a single realization (i.e., a single year) X^i, remembering that each x_t^i has $M * H$ dimensions. Assume that the system has a finite number N_s of different and physically significant states s:

$$S_t = \{s,\, 1 \leq s \leq N_s\} \tag{2.27}$$

We can thus define a state sequence S_t^i which contains the states s_t^i of each x_t^i, mapped by a function δx_t^i. Hence the mapping of matrix D_f results in matrix D_s which is a matrix of the states:

$$D_f = \begin{bmatrix} s_1^1 & \dots & s_1^T \\ \vdots & \ddots & \vdots \\ s_{365}^1 & \dots & s_{365}^T \end{bmatrix} \tag{2.28}$$

This mapping, at the light of what we said before, is performed on the base of seasonal cycle and residual component of each row.

Now the recognition of seasons lies in the definition of the mapping function. As we will see in the next chapter, the aim of the machine learning approach is to implicitly recognize the mapping function which best catches this path without human supervision, and in this way defines the meteorological seasons.

Chapter 3

Data

There is currently a large number of climate datasets available within the climate community, developed and distributed for a wide range of uses. They differ for several characteristics, such as the space and time coverage and resolution, and the variables which they provide. In this work we make use of two kinds of products, namely one climate reanalysis and climate models providing historical and future simulations. This chapter presents the used datasets – the ECMWF ERA5 reanalysis [Hersbach et al., 2020] (section 3.1) and the Earth System Model EC-Earth3 [Döscher et al., 2022] (section 3.2), along with the physical variables considered in this study, the total precipitation and surface air temperature.

3.1 Climate reanalysis: ERA5

3.1.1 ERA5

ERA5 is the fifth generation climate reanalysis developed and distributed by the European Centre for Medium-Range Weather Forecasts (ECMWF). Climate reanalyses are datasets that combine, through data assimilation techniques, historical observations with the output of numerical models to provide a detailed gridded picture of the past and present state of the climate system on a global scale at the surface and for all levels of the atmosphere. Reanalysis data have been widely applied in atmospheric sciences, for example, to assess the impact of changes in observing systems or to compute state-of-the-art climatologies [Hersbach et al., 2020]. In this work, ERA5 is used as a ground-truth dataset, that is to say the dataset which contains the truth and is therefore used to evaluate the performance of the built methodology.

3.1.2 Reanalyses vs observations

What makes a climate reanalysis system appreciated and reliable is the fact that it is able to assimilate observations into one physical-dynamical model. Observations would be the best possible source of information to understand the current and recent past climate, however they are characterised by a number of drawbacks and weak points, including:

- their spatial and temporal domain: observation datasets provide information only at specific times and locations, corresponding to the station's operative periods and locations. In-situ stations are sparse and unevenly distributed over the globe (e.g. valleys vs mountains; land areas vs sea). More regular observation datasets, such as the ones obtained from satellite data, only cover the more recent period (typically the last 40 years, from 1979 on).

- their consistency: observational data, being obtained by different sources, may have some bias between each other and therefore they have to be verified and homogenized before their use.

- their accessibility: there is a large number of observation datasets available, but not all of them are easily accessible.

A climatic reanalysis can overcome these issues by processing the observation data within a physical model, creating homogeneous space-time grids of standardized and verified data. ERA5 incorporates data from a great number of observation sources which are assimilated, processed with the physical model IFS Cy41r2, and then stored into hourly fields. More information about the ERA5 model workflow and settings could be found in the reference paper [Hersbach et al., 2020].

3.1.3 ERA5 data

In this work we will use the two following ERA5 variables: "total precipitation" and "surface air temperature". Even if ERA5 data are available from 1950, we selected only the period from 1979 to 2020, as the data in this period are considered more reliable as they assimilate also satellite observations, which are available starting 1979.

ERA5 has a spatial resolution of 0.25°x0.25° (about 30 Km). Both variables are used in their daily mean temporal aggregation.

ERA5 dataset is distributed through the Copernicus Climate Data Store (CDS) portal [ECMWF, 2023].

3.2 Climate simulations: EC-Earth3

3.2.1 EC-Earth3

EC-Earth3 is a Earth System Model (ESM). ESMs are a class of numerical models which aim to describe the behavior of Earth's climate system. ESMs are the upgrade of Global Climate Models (GCMs). GCMs have been designed as a combination of coupled models describing the atmosphere, sea ice, ocean, and land. ESMs also include components for vegetation and carbon cycle in order to perform simulations more representative of the behavior of the entire Earth's climate system. Nowadays, ECMs are our best tool to understand the Earth's climate system and its possible future evolution. In this work we will use the basic configuration of EC-Earth3, which includes the components describing atmosphere, sea ice, ocean, and land.

EC-Earth3 is developed on the concept of "seamless prediction". That is to say that a seasonal weather forecast model, which simulates atmospheric dynamics and thermodynamics over short time scales, is joined with a climate model, which simulates the interactions between the atmosphere, ocean, land surface, and ice over longer time scales.

The models used for each components are [Döscher et al., 2022]:

- IFSr4, developed by ECMWF for the atmosphere module, with a horizontal resolution of about 80 km and 91 vertical levels. It includes the land model HTESSEL.

- NEMO 3.6, developed by the Nucleus for European Modelling of the Ocean, for Oceans, with an average horizontal resolution of 1° x 1° and 75 vertical levels.

- LIM3 for sea ice model developed by Louvain la Neuve.

These modules are coupled through The OASIS3-MCT coupler version 3.0. For more details on the setup of EC-Eart3 refer to [Döscher et al., 2022].

In this work we make use of the simulations of the EC-Earth3 model performed for contribution to the Coupled Model Intercomparison Project Phase 6 (CMIP6). CMIP6 collects the results of over 100 models from more than 50 modeling centers around the world. In the next subsections we will give a brief overview of the design of CMIP6 experiments.

3.2.2 CMIP6

The Coupled Model Intercomparison Project (CMIP) is now one of the foundational elements of climate sciences. CMIP started over 20 years ago as a comparison of the first global coupled climate models (numerical physical models which simulate different components of the Earth system and their interaction) and now has reached its 6th phase (CMIP6). CMIP gives the baseline for the model settings and collects and distributes the outputs obtained by models developed by more than 50 modeling centers around the world. Due to the increase of the scientific questions that these models try to answer, along with the increase of information that these models could give, CMIP6 has reorganized its structure with respect to the precedent phase (CMIP5). Now three major components could be identified, as detailed by [Eyring et al., 2016]:

- the Diagnostic, Evaluation and Characterization of Klima (DECK) experiments (klima is Greek for "climate"), and CMIP historical simulations. DECK includes four baselines simulations: 1) an historical Atmospheric Model Intercomparison Project (amip) simulation, 2) a pre-industrial control simulation (piControl), 3) a simulation forced by an abrupt quadrupling of CO2 (abrupt $4\times CO_2$) and 4) a simulation forced by a 1 % yr^{-1} CO_2 increase (1pctCO_2). The historical simulation is designed to cover the recent past period (1850-2014) (section 3.2.2). These simulations are essential because they provide a standardized baseline for model comparison.

- The creation of a common infrastructure with standardized documentation which facilitates the distribution of the models results.

- The reorganization of the experiment runned for the project, called CMIP6-Endorsed Model Intercomparison Projects (Endorsed MIPs), which led to the creation of the guidelines for 23 specific research projects.

In this work we will make use of the historical simulation and of the Scenario Model Intercomparison Project (Scenario MIP), which is the Endorsed MIP designed to evaluate the response of climate models to different future emissions and socio economics scenarios.

3.2.3 Historical and future scenarios simulations

The historical experiment is a simulation of the recent past (1850-2014), in which changing conditions are imposed consistently with observations. The guidelines for the execution of the experiments requires at least one ensemble member and the

use of a Atmosphere-Ocean coupled general circulation model. For all the forcing constraints, proper datasets are indicated by CMIP6. For a complete reference see [Documentation, 2018]. The rationale behind historical experiment is to evaluate the models performance against present climate and observed past climate changes.

The Scenario Model Intercomparison Project (Scenario MIP) [O'Neill et al., 2016] is the primary activity in CMIP6 which provides climate projections based on alternative scenarios of future emissions and land use changes and has been designed with eight alternative 21st century scenarios. These scenarios describe the possible future developments of anthropogenic drivers of climate change. Until CMIP5 these scenarios consisted of Representative Concentration Pathways (RCPs), a set of four pathways of land use and emission of air pollutants and greenhouse gasses. In CMIP6, these pathways have been integrated with the Shared Socioeconomic Pathways (SSPs), which modelize socioeconomic development. The idea behind this choice is to focus not only on the physical climate system, but also on the climate impacts on societies. RCPs are named after the radiative forcing (the balance alteration between incoming and outcoming energy in the Earth system) they produce in 2100, measured in W/m^{-2}. CMIP6 incorporates seven RCPs: 1.9, 2.6, 3.4, 4.5, 6.0, 7.0 and 8.5. SSPs are organized into 5 levels: SSP1 and SSP5 envision optimistic trends for human development but, while SSP1 assumes a shift toward sustainable practices, in SSP5 there is an energy intensive, fossil based economy. SSPs 3 and 4 envision more pessimistic development trends, with increasing inequalities. SSP2 prospect is a central way in which trends continue their historical patterns. Figure 3.1 summarizes the SSP-RCP scenarios used in CMIP6.

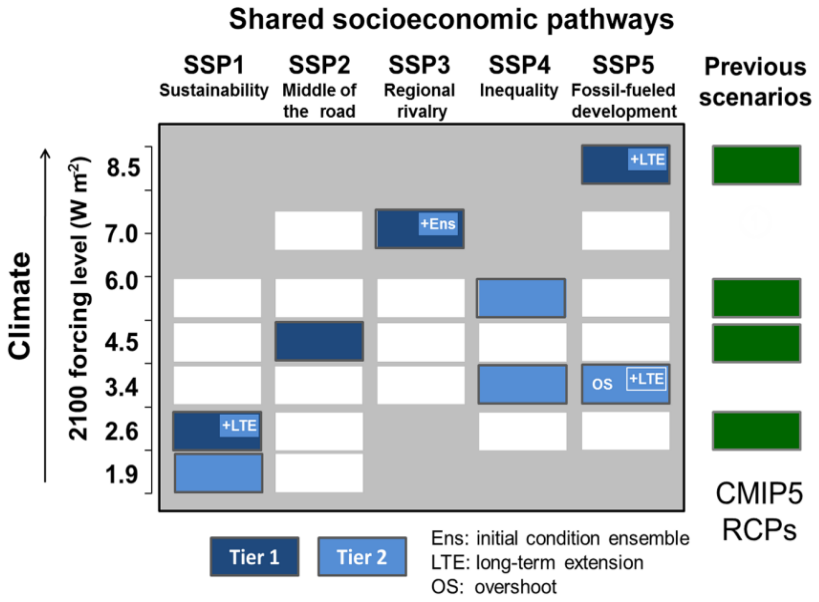

Figure 3.1: **SSP-RCP scenario matrix** illustrating ScenarioMIP simulations. Each cell in the matrix indicates a combination of socioeconomic development pathway (i.e., an SSP) and climate outcome based on a particular forcing pathway (i.e., an RCP). Dark blue cells indicate scenarios that will serve as the basis for climate model projections in Tier 1 of ScenarioMIP; light blue cells indicate scenarios in Tier 2. White cells indicate scenarios for which climate information is intended to come from the SSP scenario to be simulated for that row. CMIP5 RCPs, which were developed from previous socioeconomic scenarios rather than SSPs, are shown for comparison (Source [O'Neill et al., 2016]).

3.2.4 Enseble members

The simulations detailed in the previous section are performed by each model participating in CMIP6. The result is the so-called "multi-model ensemble". The analysis of the multi-model ensemble is primarily used to explore the spectrum of possible evolution of Earth's climate system under the conditions imposed in the specific experiment.

Another approach carried in CMIP6 is the so-called "multi-member ensemble". Each model performs the simulations slightly varying the experiment setup to obtain a spectrum of results for each experiment. Taking as example EC-Earth3 models, the historical simulation is carried out several times changing the setup, obtaining different historical simulations which are the multi-member ensemble of EC-Earth3 model for historical period. The same approach is used in the ScenarioMIP experiments and in the other EndorsedMIPs. The rationale behind multi-member ensembles is to use this spectrum for the evaluation of the model's sensitivity to a slight change in the setup.

In CMIP6, each member of the multi-member ensemble is identified with an univocal code, called "VARIANT-ID". VARIANT-ID are encoded in the form $r[r_{idx}]i[i_{idx}]p[p_{idx}]f[f_{idx}]$, where each index is an integer (≥ 1) and corresponds to

[Taylor et al., 2018]:

- r_{idx}: the realization index, used for distinguishing among members of an ensemble of simulations that differ only in their initial conditions. Each future scenario simulation should be assigned the same realization integer as the historical run from which it was initiated.
- i_{idx}: the initialization index, used either to distinguish between different algorithms used to impose initial conditions on a forecast or to distinguish between different observational datasets used to initialize a forecast.
- p_{idx}: the physics index used for identifying the physics version used by the model.
- f_{idx}: the forcing index, used to distinguish runs with different variants of forcing applied.

Normally, for the multi-model ensemble, only an ensemble member for each model is used, usually the r1i1p1f1 member.

3.2.5 EC-Earth3 specifics

As for ERA5, we will use the "total precipitation" and "surface air temperature" variables. We will make use of the Historical simulation (1850-2014) and of the future projection under the SSP5-8.5 from ScenarioMIP (2015-2100). The historical simulation partially overlaps with ERA5 (1979-220) and will be used for the comparison between the two datasets, such as for the evaluation in the past of the methodology we developed.

In this work, dealing with only one model (EC-Earth3), we will use a multi-member ensemble. We selected the VARIANT-ID with both total precipitation and surface air temperature available for both historical simulation and future projection under SSP5-8.5 scenario. These requirements lead to the identification of three ensemble members (r1i1p1f1, r13i1p1f1, r15i1p1f1). EC-Earth3 has a spatial resolution of 0.70°x0.70° (about 80 Km). Both variables are used in their daily mean temporal aggregation.

EC-Earth3 dataset, such as the whole CMIP6, is available on the portal of the Earth System Grid Federation (ESGF) [ESGF, 2023].

Chapter 4

Methods

This chapter aims to introduce the development of a method to objectively identify seasons in climate data, aiming at being flexible and highly adaptable to different case studies. Thus, the choice of a data driven approach appears to be the most obvious consequence. In the last 30 years data driven methods, commonly defined as "machine learning", have been the subject of great interest from the scientific community. The continuously increasing amount of available data has given rise to the need of methods which can extract and condense relevant information with as little human interaction as possible. In this direction machine learning has given multiple proofs of being able to achieve this task. Climate sciences have faced a similar issue, given the enormous increase of data availability in the last decades, and machine learning is yielding promising results in this field too.

This chapter presents the methodology developed in this work, which has been built in the light of being adaptable to the most diverse cases. Section 4.1 contains an overview on machine learning methods and their history, with a special focus to those concerning climate sciences. In sections 4.3 and 4.4 are respectively presented the method adopted for the seasons definition, and the one for the seasons projection.

4.1 Machine learning

Note to the reader: an exhaustive review of machine learning history and methods goes beyond the goal of the current thesis. The purpose of this section is to point out the main ideas and evolution of the methods which inspires this work.

4.1.1 History

The term Machine Learning (ML) refers to a wide class of algorithms and statistical models which aim to perform specific tasks without being explicitly programmed for them. The origin of ML can be placed between the end of the 50s and the beginning of the 60s of the XX century, when Rosenblatt performed the first mathematical studies about the perceptron [Rosenblatt, 1959], with the task of making a machine recognize some hand-written numbers. The perceptron is a binary classifier which maps input values in output classes with:

$$f(x) = \chi(< w, x > +b) \tag{4.1}$$

Where x is the input data, w the so-called weights, b the bias and $<,>$ denotes the internal product. In Rosenblatt formulation was a threshold function which gave 1 if $< w, x >$ is bigger than b and 0 otherwise. Both the parameters w and b are optimized in the training process, where they are randomly initialized and then corrected on the prediction they give on the data. In this sense, a real breakpoint in ML history was the proof given by Novikoff of the learning algorithm convergence [Novikoff, 1962]. In 1962 Widrow created MADELINE, a perceptron with an additional layer between the input and the output called hidden layer, giving birth to the first multilayer neural network [Widrow and Stearns, 1990].

Formally, the training process of a perceptron could be seen as the identification of a hyperplane in the phase space which properly divides the features of the data based on their belonging class. This problem was firstly approached deterministically, since Tsypkin in 1968 showed the power of stochastic methods [Tsypkin, 1968].Taking as reference the formula 4.1 for the perceptron, he introduced a performance index $J(w, x)$ as the expectation value E_x of a generical cost function $Q(w, x)$, called loss function,which quantifies how the current weights w allows a correct identification of the real belonging class of the data x:

$$J(x) = E_x Q(w, x) \tag{4.2}$$

Thus, the goal of the learning process is to minimize $J(w, x)$, and Tsypkin himself proposed a learning algorithm known as Stochastic Gradient Descent (SGD):

$$w[n] = w[n-1] - \gamma[n] \nabla Q(x[n], w[n-1]) \tag{4.3}$$

Where $\gamma[n]$ represents the rate at which the weights w are updated.

In this period also the first clustering algorithms were developed, such as K-means clustering [Lloyd, 1957] and Hierarchical clustering [Ward Jr and Hooker, 1963], with the purpose of grouping data in clusters based on their characteristics. In 1969 a book by Minsky and Papert showed some limitations of the perceptron [Minsky and Papert, 1969], driving the beginning of the so-called 1st ML Winter in which development of ML was quite limited. A new breakthrough in ML advances was the introduction of the back-propagation algorithm, which updates the weight of each layer starting from the last one with a chain rule, instead of updating all of them at the same time [Rumelhart et al., 1986]. Despite a new period without remarkable achievement (commonly referred as the ML 2nd winter), since the 1990s ML has experienced a new boom. Three pushing factors could be recognized to explain this new phase:

1. The continuously increasing amount of data, which makes the extrapolation of information from them more a necessity than a scientific curiosity.
2. The decrease in parallel computing and memory cost.
3. The development of new machine learning algorithms.

4.1.2 Most used machine learning techniques

ML is a wide and multidisciplinary sector, and relies on a great variety of algorithms, which are applied depending on the specific task. A survey in these methods could recognize three main categories: supervised learning, unsupervised learning, and reinforcement learning [Mahesh, 2018].

- **Supervised learning**: the purpose of supervised learning is to instruct a computer system to predict output values of a system based on a set of input values. This prediction could be either of the class to which the data belongs (classification), or one or more continuous variables (regression). Thus, a supervised algorithm needs to be trained on a labeled dataset, i.e., a dataset where each input data is associated to its belonging class or output value, and can only be used on other data once trained on this initial labeled dataset.
- **Unsupervised learning**: unsupervised learning aims to extrapolate relationships from complex data without relying on labeled data. This can be achieved by grouping data according to their characteristics (clustering), or by determining the data distribution (density estimation), or even reducing the dimensionality of data (Principal Components Analysis).

- **Reinforcement learning**: reinforcement learning is the ML technique which better reproduces the human learning process. Here the algorithm learns the best behavior by a sequence of states and actions with a system of reward based on the choices taken. Reinforcement learning is widely used in games and other fields that involve human interaction.

A cross-sectional area to these categories is the Neural Networks (NNs) field. NNs, also called Artificial Neural Networks (ANNs), are computer systems which try to emulate the simplified model of a biological neural network. The constituent unit of a NN are the artificial neurons, which are interconnected nodes organized in layers. The shape of these layers and the type of connections (i.e., the architecture of the NN) could vary considerably depending on the purpose for which it is being implemented, and a comprehensive review of all their applications is beyond the scope of the current thesis. Nevertheless, a common base structure in NNs is formed from three layers: an input layer, a hidden layer, and an output layer. Due to the presence of multiple layers of representation, neural networks are an example of the so-called Deep Learning. NNs have been proved to be well performing in a wide class of applications, especially when dealing with nonlinear problems. Furthermore, by manipulating their architecture, they could be used for different purposes such as classification, regression, dimensionality reduction and reinforcement learning.

4.1.3 Machine learning in climate sciences

As many other sectors, climate sciences have experienced an extraordinary increase in data availability (Figure 4.1). Consequently, it can be considered as being an example of the so-called big data, defined by their 'four Vs': volume, velocity, variety, and veracity (Figure 4.2). These features make the data hard to manage. On the other hand, it is now clear that addressing climate changes involves adaptations (preparing for the inevitable consequences), and this data is the core of the strategies that can be implemented. This is forcing the scientific community to face the problem of climate data diffusion and interpretation, in the light that the information that these data contain must be transposed in an easily accessible form for policy makers [Overpeck et al., 2011].

For these reasons, machine learning algorithms are being used with increasing frequency in the field of climate data and sometimes they can provide better results than "more" classical statistics models. Nevertheless, a massive implementation of ML on Earth system data is still lacking, and the data analyst community is trying to give itself guidelines to fill this gap. Some of the fields in which ML is giving better results are [Reichstein et al., 2019]:

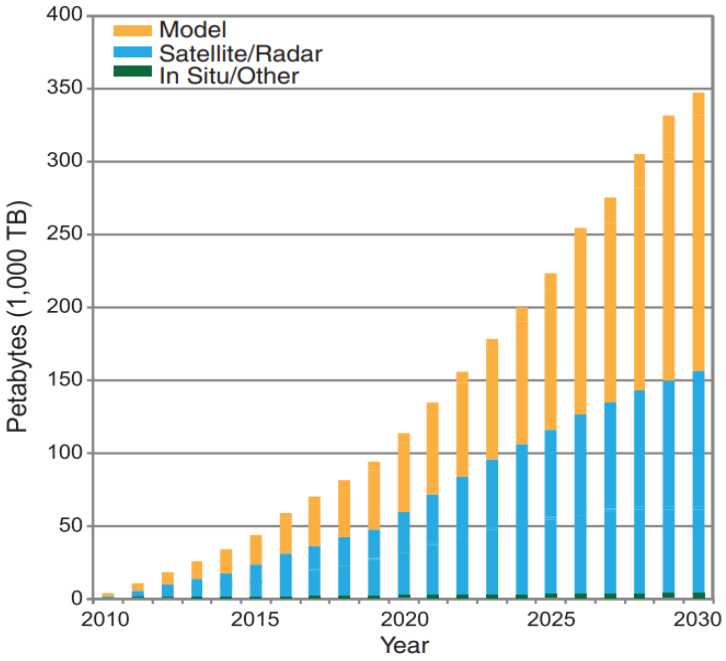

Figure 4.1: **Estimation of the volume of climate data**: (source of image [Overpeck et al., 2011])

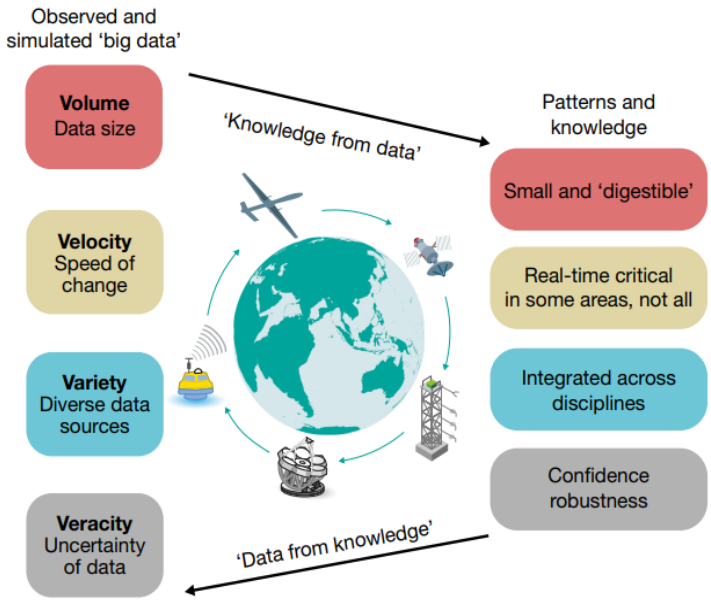

Figure 4.2: **The 4 Vs of earth system data** (left) and the main features that should came from their analysis (right) (source of image [Reichstein et al., 2019])

- **Global modeling**: ML is finding its application supporting numerical simulations as those provided by Global Circulation Models: a new sector of ML is emerging aiming at developing algorithms able to learn the behavior of dynamical systems, such as the earth system, making use of different techniques. The most used methods are Physics-Informed Neural Networks (PINNs) and Neural Ordinary Differential Equations (Neural ODEs). PINNs are NNs designed to include the governing equations and constraints that describe the system being modeled, making the network able to learn the underlying physics of the system and make predictions that are consistent with the laws of physics [Cai et al., 2021]. PINNs are particularly useful to speed up and optimize GCMs execution. Neural ODEs are a type of neural network architecture that allows for the modeling of dynamic systems as a continuous-time differential equation. In a Neural ODE, the input is a set of initial conditions for the system, and the output is the state of the system at a future time. The network learns a set of continuous-time differential equations that describe the evolution of the system. This means that Neural ODEs can be used to extrapolate the behavior of a system beyond the range of available data [Chen et al., 2018]. This is particularly useful for fields that have a large availability of data and a lack of reliable physical relationship (for example ice sheets dynamics).

- **Downscaling**: for the analysis of future scenarios, most of the time it is necessary to transpose the output global fields obtained with GCMs into projection over specified locations. This operation, called downscaling, is performed with a great variety of dynamical and statistical models. Dynamical downscaling is performed using Regional Climate Models (RCMs) which make use of the outputs of GCMs as boundary conditions. ML could improve the performance of these models with the methods explained above. Statistical downscaling is performed establishing the relationship between GCMs outputs and variable behavior at local scale. Here, ML has proved to give reliable results with the application of recurrent and convolutional neural networks, that is to say neural networks which can take into account the space and time features of the dataset [Reichstein et al., 2019].

- **Clustering**: notable results have also been reached with the application of clustering algorithms to climate data, with a wide range of goals. Clustering can be used to group together similar regions or time periods based on climate variables, such as temperature or precipitation. This can help identify regions that are particularly vulnerable to climate change, or identify trends and patterns that may be related to global climate phenomena. Clustering, being an

unsupervised method, allows to overcome some classicals methods which rely on heuristic decision rules. Many studies have attempted to perform climate classification from either a global and regional point of view, querying the number and the nature of the variables that should be included in such analysis. Also, clustering techniques have been applied to improve the computation of some climate indexes, with a better recognition of the spatio-temporal domain in which they are defined [Steinbach et al., 2006].

4.2 Data preprocessing

Before proceeding with the discussion of the methods used in this work for the division into meteorological seasons, we describe the preprocessing of the data, i.e, the operations which make the datasets suitable for our analysis.

Climate datasets, such as ERA5 and Ec-Earth3, could be visualized as five-dimensional tensors. The first dimension represents the physical variables (in our case, surface air temperature and total precipitation), the second, third and fourth dimensions represent the spatial coordinates (respectively, longitude, latitude and vertical level), while the fifth dimension represents the time coordinate. The variables used in this work, surface air temperature and total precipitation, are distributed on single levels. This means that they do not need the vertical level coordinate, since they are evaluated only near the ground (surface air temperature) or on the ground (total precipitation). Thus, the datasets used in this work could be visualized in four-dimensional tensors where each dimensions represents respectively the physical variables, the longitude coordinates, the latitude coordinates, and the time coordinates.

4.2.1 Data remapping

For the application of the methods we will describe, it is necessary that the data tensors we will use (one for ERA5 and one for EC-Earth3) share the same shape. As detailed in chapter Data, for both datasets we will use surface air temperature and total precipitation in their daily mean time aggregation. Thus, the first and the fourth dimensions are consistent. On the other hand, the space coordinates are not consistent since ERA5 has an horizontal resolution of $0.25° \, X \, 0.25°$ and EC-Earth3 has an horizontal resolution of $0.70° \, X \, 0.70°$. Thus, we must remap the two tensors in the same horizontal grid.

In this work we decided to remap the ERA5 dataset in the EC-Earth3 grid. That is to say, we standardized both datasets on the coarsest spatial grid. Remapping on

the finest grid (tecnically called downscaling) is a legit operation as well, but would require more caution and the application of specifics methods, since it implies generating information at a resolution which is not the one of the original datasets. We remapped ERA5 on EC-Eart3 grid using the conservative interpolation. Formally, any value obtained with interpolation could be written in the form:

$$\overline{f} = \sum_{\sigma} f_{\sigma} w_{\sigma} \tag{4.4}$$

Where $\overline{f}$ is the value in the interpolated field, σ tags the elements f in the original field that contribute to interpolation, and w are the interpolation weights. In conservative interpolation, σ tags the original grid cell which overlap with the resulting grid cell, and w is the ratio of the area shared by the original and the resulting grid cells (ref)(Fig 4.3).

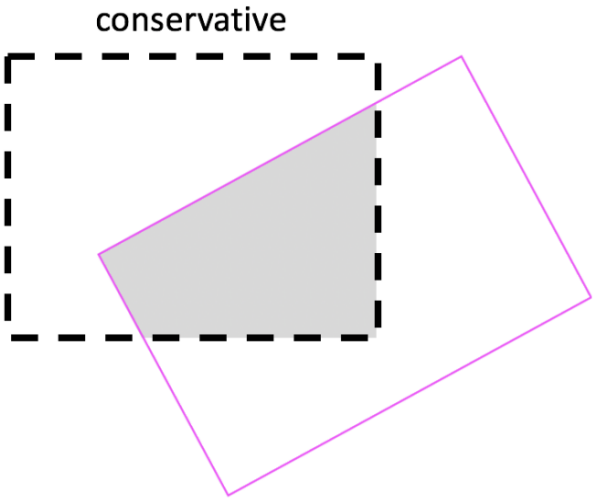

Figure 4.3: **Calculation of conservative interpolation** weights w for a original grid cell (dashed lines). Violet lines represent the area covered by the resulting grid cell. The weight associated with the resulting cell is the ratio of the shaded area over the original cell area. (Source of image: [Pletzer and Hayek, 2018])

Conservative interpolation is particularly indicated for the total precipitation field, since it allows to conserve the spatial total amount of precipitation. That is to say, if we select an area in the original field and the same area in the interpolated field, the total amount of precipitation is the same.

4.2.2 Moving averages

The second operation performed in the preprocessing phase relies on some heuristic considerations. Climate time series are subjected to variability at high frequencies, caused by the complex mutual interactions between the components of Earth's climate system. Relying on the simplified recognition we made of time series component in chapter Seasons and Seasonality, we can include this variability in the residual part. Since our purpose is to recognize the seasonal patterns, we assumed that this variability could lead to results that are more difficult to interpret. As we will see in section 4.4, a part of our methodology will be focused on recognizing the the season to which each day belongs. Thus we can suppose that a high variability between days could comport high variability in the results.

We tried to reduce this variability applying a moving average μ_{mov} to the original data. The moving average is a commonly used operator for the empirical reduction of high frequencies variability in time series analysis. Given a time series $X = \{x_1, \ldots, x_N\}$, the moving average is computed for each value x_n as:

$$\mu_{mov}[x_n] = \frac{1}{k} \sum_{i=n-(k/2)}^{n+(k/2)} n_i \tag{4.5}$$

Where k is the amplitude of the so-called window of the moving average.

Firstly, we applied the moving average on time dimension for each variable and grid point, with a window of $30\,days$. Then, in order to strength more the seasonal signal, we applyied the moving average through year for each ordinal day of the year, with a window of $30\,yr$. That is to say, considering for example the January 1^{st} of each year, we computed the moving average through January $1^{st}s$.

In this way, each day in our datasets now contains information of the previous and following 15 *days* and *yr*. Furthermore, both intra-annual variability and inter-annual variability have been empirically reduced.

Since we computed moving average only on complete windows, this operation removes the firsts and lasts 15 *days* and *yr* in the datasets. It's worth noting that this operation is not necessary and has been performed only to obtain more clear ad interpretable results.

4.3 Clustering: a Radially Constrained method

The main goal we want to achieve is to build a method for the definition of meteorological seasons. In light of the previous section, unsupervised learning has been identified as the most suitable choice since it allows us to extract information from the data without any a-priori assumption. More specifically, grouping climate data into seasons seems to be a task suitable for clustering. Due to their extensive use in many different applications, a wide number of clustering algorithms have been developed. Nevertheless, all of them share the same purpose, which is to group the input in order to [Xu and Wunsch, 2015]:

- Maximize similarity between items in the same cluster.
- Maximize the differences between items in different clusters.
- Perform the previous operations based on a metric which is descriptive of the dataset and fits for the purpose of clustering.

The standard procedure in the development of a clustering method consists of [Xu and Wunsch, 2015]:

- Extract the most relevant feature from the dataset according to the purpose of the work.
- Design the algorithm in order to catch these features in a proper way.
- Evaluate the performance of the algorithm.
- Explain the obtained results.

In the rest of this section, we will follow this workflow.

4.3.1 Features extraction

The theoretical basis of the seasonal feature extraction from the dataset is detailed in chapter two, and below is reported the operative process. The input data consists of a four-dimensional matrix from a gridded dataset (with shape longitude pts X latitude pts X time steps X variables) where time steps=365*years being daily the sampling frequency. Data are reshaped in a two-dimensional matrix (365 X (years * latitude pts * longitude pts * variables)), to obtain a representation of data where each day is an item (i.e. the object that will be assigned to a cluster) and has as features the values each variable had in this day for each year and each grid point.

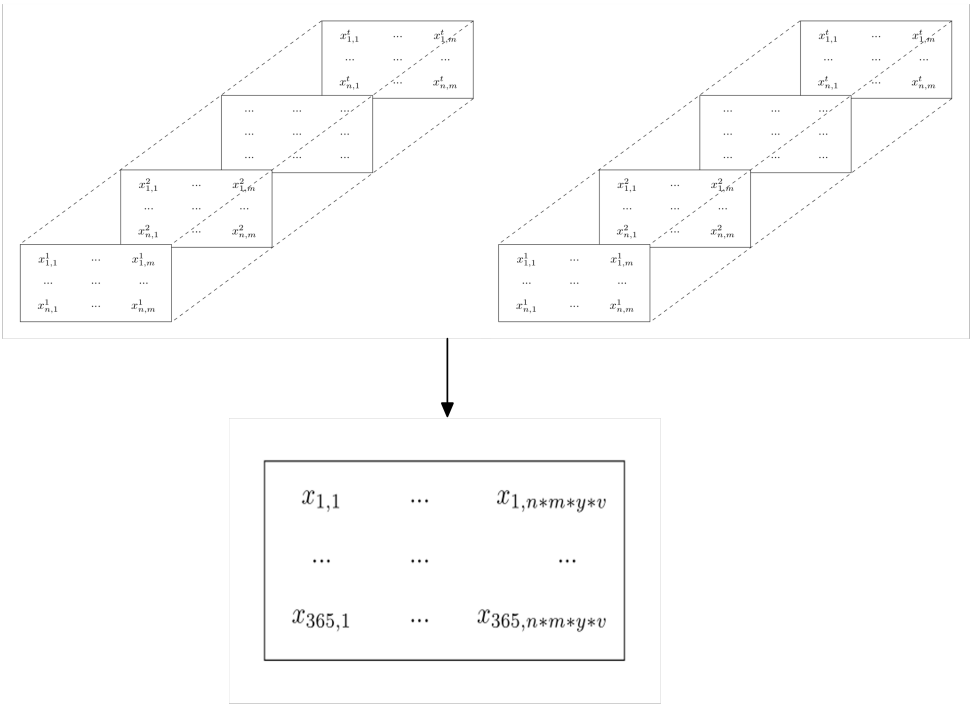

Figure 4.4: **Schematic of data reshaping**: n= longitude points, m = latitude points, t = time steps, y = years, v = variables)

In order that different dimensionalities and variance do not affect our detection, the features used for the clustering must be scaled in order to be comparable. Eventual inter-year trends and differences in space and physical variables absolute value may force the algorithm to perform the cluster based on information that is not relevant to the seasonal cycle. Therefore, the features are scaled with standardization, (or z-score normalization), so for each column X_j in the matrix obtained in Figure (4.4):

$$X_j = \frac{X_j - \mu[X_j]}{\sigma[X_j]} \tag{4.6}$$

Where μ is the mean and σ the standard deviation. The data representation obtained is used for the clustering.

4.3.2 Algorithm design

In this work we used an algorithm inspired by the paper *Defining climatological seasons using radially constrained clustering* [Cannon, 2005]. Cannon proposed an algorithm which could be placed in the class of the clustering algorithm based on partition. Such methods aim to perform a classification of the data into a set of

disjoint clusters, based on a specific metric, in order to reach the purpose introduced at the beginning of this chapter [Xu and Wunsch, 2015]. In this way, the result of the clustering provides both the grouped data and a statistic performed on the partitions obtained on the features space. A widely known algorithm from this class is the k-mean which defines the partitions based on the means of the values of the item contained in them (the so-called cluster centroids). The main workflow for this class of clustering algorithm could be synthetized as done by [Xu and Wunsch, 2015]:

- Generate random centroids,
- Compute the metric,
- Update centroids with a defined method,
- Repeat the previous step until the metric converges.

The more evident problem in the application of these kinds of algorithms to our case is the management of time dimension. The definition of meteorological seasons implies that the resulting clusters are time-contiguous, but our dataset does not contain any explicit information about the time location (i.e. the day of the year) of each item. A possible solution could be achieved by introducing one or more fictitious features (for example, adding a dimension which expresses the day of the year), but we discarded this option to avoid improper conditioning of the problem that would artificially drive the cluster solution. The second problem, still related to time dimension, is that the periodicity of the dataset, i.e. that contiguity must be respected on the boundaries of the dataset, so that the last element (corresponding to December the 31^{st}) is contiguous to the first one (January the 1^{st}). In other words, the cyclicity of the seasons must be respected.

Radially constrained clustering algorithm allows to overcome these problems, forcing the clusters to be time contiguous and assuming the correct periodicity. Practically, this is achieved by defining - instead of the centroids as done by k-means - the time breakpoints which divide the clusters.

To provide a more concrete example, please consider the dataset obtained in Figure 4.4 made by 365 samples of D dimension x_t^d where $t = 1, \ldots, N$ represent the time and $d = 1, \ldots, D$ the dimension, where $D = n * m * y * v$. Data must be ordered over time, which means that xi and x_{i+1} are time contiguous. Furthermore, data must be periodical, so contiguity must be respected also for x_{365} and x_1. It must be noticed that this continuity is not strictly respected, since for each column the first and the last element are not contiguous. Nevertheless, for each variable and grid point, each column is contiguous to the following one. We can thus suppose that contiguity is broadly respected if the number of years is large enough to absorb the

information at boundaries. We set a limit of 30 years, which is the commonly used time window for climate analysis. The aim is to determine M clusters by defining M temporal breakpoints b_k with $k = 1, \ldots, M$. The goal of the algorithm is to minimize the Within Sum of Squares (WSS):

$$WSS = \sum_{M} \sum_{N} \sum_{D} (x_i^d - \mu_i^d)^2 \tag{4.7}$$

This metric, which is an euclidean distance. has been chosen as it is the same proposed by Cannon. Furthermore, euclidean distance is the most used metric in clustering algorithms. A future development of this work could consider a systematic comparison between different metrics. The algorithm proposed by Cannon does not have an implementation, so part of this work consisted of the practical realization of it. The core of the algorithm consists of the following steps:

1. Starting breakpoints are randomly generated and are bound to be equally time spaced.
2. WSS is computed.
3. Breakpoints are updated, each of them adding a random integer number $u_k \in U(-L, +L)$.
4. WSS is computed again, if smaller than WSS of previous step, new breakpoints are accepted, otherwise breakpoints are downgraded to the previous version.
5. Steps 3 and 4 are repeated until WSS converges to its minimum.

To improve the algorithm, the following optional are added:

1. A scheduler for update rate, which scales down L if the metrics are getting smaller at a very slow pace.
2. A constraint on season length: if an iteration violates it, the previous breakpoints are restored.
3. An ensemble method, which performs the clustering several times with different starting breakpoints and then keeps the best results according to WSS.

As for K-Means, this algorithm has a weak point in the definition of M, the number of clusters, which in our case correspond to the number of seasons. There is not a general and objective way to define this hyperparameter, but some criteria which can be used to evaluate the goodness of a certain choice. The following subsection contains the criteria chosen in this work.

4.3.3 Evaluation metrics

The evaluation metrics in this work have a triple goal:

1. Evaluate the reliability of the clustering algorithm,
2. Evaluate the best number of clusters to be used to divide the dataset (i.e. the best number of season)
3. Give an answer on the work hypothesis we made in chapter 2, i.e. if the clustering approach is suitable for the seasonal division.

The elbow method is a qualitative way to evaluate the optimal number of clusters in a dataset, based on a plot called "elbow graph". The clustering is performed with a various number of clusters, and then for each run the WSS at convergence is plotted. Ideally, the WSS decreases when the number of clusters increases, and the rate of this decrease is called "gain". The ideal number of clusters is chosen as the one after which the gain decreases and is recognizable in the plot due to the characteristic elbow shape [Yuan and Yang, 2019]. This method, albeit qualitative, is commonly used for the evaluation of the most proper number of clusters. On the other hand, the silhouette score is defined to compare the similarity between data in the same clusters with differences between data in different clusters. The silhoutte coefficient is defined, for each element xi as:

$$s_i = \frac{b_i - a_i}{max\{b_i, a_i\}} \tag{4.8}$$

Where ai is the average distance between xi and the other element in the same cluster, and bi the average distance between xi and the elements in the other clusters. This coefficient is in the range [-1,1] and approaches 1 when there is a close relationship between the object and the assigned cluster [Yuan and Yang, 2019]. The silhouette score is obtained averaging that coefficient over all the data, and tested along a various number of clusters. Thus, the optimal number of clusters is the one which maximizes the silhouette score. This method is also used for the evaluation of the algorithm: low values indicate a general bad performance .

4.3.4 Results interpretations

The radially constrained clustering algorithm gives M temporal contiguous clusters, where M is the optimal number defined on the criteria exposed in subsection 4.3.3.

Being the dataset daily, this means that each day is assigned to a cluster. These clusters are the data-driven defined meteorological seasons. The following step is to study the evolution of the seasons in future climate projection, which corresponds to tracking the evolution of these clusters in new data.

4.4 Seasons projection: the SoftMax perceptron

The clusters obtained are defined by the temporal breakpoints, so a reasonable approach to investigate the evolution of seasons to future projections in climate model data could in principle be achieved by applying these breakpoints to the new data, and then study the evolution of the physical values in the new clusters. However, this approach has been discarded, since our purpose is to study how the current definition of seasons will evolve, and this could imply a variation in their onset and withdrawal, which cannot be captured with this method.

Another way analysis of the seasons in future climate could be obtained by computing the clustering on the new data, losing memory about the ones computed on historical data. This would allow us to obtain a dynamical definition of the breakpoints, but again new clusters may not be correlated to the previous ones, preventing us from exploring the evolution of the present seasons.

Considering the above points, we decided to rely on supervised learning: the obtained clusters could be used as a labeled dataset for the training, in order to make the system learn the features of the present-day seasons. Once the algorithm is trained, it should be able to assign each day of the climate projection to one of them. This method does not ensure that the resulting seasons are time contiguous. So, their eventual contiguity will be used as a criterion for the validation of the model.

The most used supervised-learning methods are the neural networks (NNs). NNs have turned out to be able to find complex structures in high-dimensional data due to their multilayered structure, and in this way establish relationships between the input data and the belonging class [LeCun et al., 2015].

Since NNs have a complex architecture which could influence the results, and therefore the architecture and the hyperparameters must be chosen and calibrated accurately, we firstly tried with a perceptron. In the case study and results chapters, we will show that the perceptron showed to be reliable , and then we assumed there is no need for a NN.

4.4.1 Perceptron architecture

As exposed in section 4.1.1 the perceptron is the progenitor of the NNs and is formed by only two layers: the input layer and the output layer (Figure 4.5). Technically, a perceptron is a binary classifier, while in this work we will face configuration where more than two seasons need to be classified. Nevertheless, the architecture of the perceptron could be generalized such that instead of estimating the probability of an event, we can estimate a vector with the probabilities of each of the multiple possible outcomes.

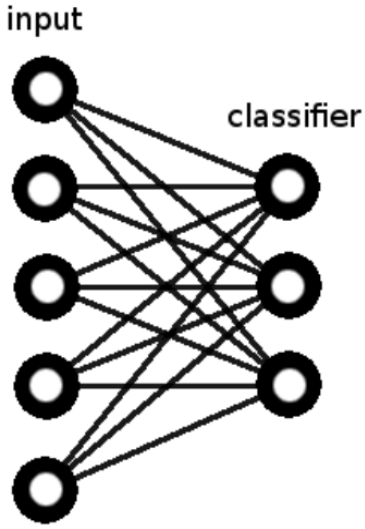

Figure 4.5: **Schematic representation of a softmax perceptron**: the lines between input and classifier units are the weights w.

Considering Figure 4.5, the input layer has N neurons, where N is the number of the features of the data, and the classifier (henceforth also called output layer) has K units, where K is the number of seasons. The perceptron computes for each unit in the output layer (i.e., for each season) the probability that the input data is associated to that unit (i.e. is associated to that season). This computation is performed with the SoftMax function [Bishop, 2006]. Being $p(y^n = k|X^n, w_k)$ the probability that the nth data is associated to the season k, we compute the SoftMax as:

$$p(X^n, w_k) = \frac{1}{\sum_{i=1}^{N} e^{w_{i,0}+\ldots+w_{i,N}}} * e^{w_{k,0}+\ldots+w_{k,N}} \tag{4.9}$$

Consequently, the sum of the probabilities for each class must be equal to 1. The training process aims to optimize the weights w_i such as the class that gets the

best score is the one that the data belongs to. This is achieved by minimizing a loss function $L(X, w)$. In this work we use the Categorical Cross-Entropy (CCE) [Lugosi and Cesa-Bianchi, 2005]. CCE is a loss function suitable for cases in which the output of the model is a probability distribution over multiple classes, as it is in our work, and is one of the most used in multi-class classification. The CCE loss function is defined as:

$$L(X, w) = -\sum_{k=1}^{K} y^n log[p(X^n, w_k)] \tag{4.10}$$

Where yn is the true class of the data. There is no analytic method for the minimization of this function, so a stochastic approach is used. The optimization of the weights is part of the training process, explained in the next subsections.

4.4.2 Dataset preparation

In our case, the dataset used for the building of the model is the one derived by the clustering. These data are currently represented by the two-dimensional matrix in Figure 4.4, with shape $(365\, X\, (\#years * \#latitude\, pts * \#longitude\, pts * \#variables))$. Each day is labeled with its season, so the labels are organized in a vector of 365 elements. A dataset of 365 elements is too small, and risks providing too few examples. So, data are reshaped in a two dimensional matrix with shape $(365 * \#years)\, X\, (\#latitude\, pts * \#longitude\, pts * \#variables)$. Thus, the labels vector is expanded by repeating itself for years times. The real values of the shape depend on the size of the region (for the $\#lat\, pts$ and $\#lon\, pts$), the included variables and the number of years in the dataset used for the seasonal clustering. These values are specified in the case study, while here are kept undefined in order to maintain generality. Once the dataset is created, it is divided into the three sets:

1. Training set (64% of total data): is used for the optimization of the weights w in the training, as will be detailed in the next subsection.

2. Validation set (16% of total data): is used for controlling the learning process. During training, the model is repeatedly evaluated on the validation set to assess its performance. It is important to note that the validation set must not be used for weights optimization.

3. Test set (20% of total data): is used to test the performance of the model in new, unseen data (section 4.4.4). For this reason, it is important that these data are not used either in training and validation processes.

This division is performed randomly, imposing that the proportion of days belonging to each season is respected in each of the aforementioned sets.

4.4.3 Learning process

The learning process aims to optimize the weights between the input and output layers such that the model could recognize the data and assign them to the correct class. The learning process consists of a pre-defined number of iterations (epochs), in which the model processes all the data in the training set. The training set is divided in batches of 128 items, then in each epoch the model processes sequentially the data in each batch. For each batch the weights are optimized in order to lower the loss. In this works we use stochastic gradient descent (SGD) optimizer, which updates the weight at each iteration n of the learning process with:

$$w[n] = w[n-1] - \gamma[n]\nabla Q(x[n], w[n-1]) \tag{4.11}$$

where is the learning rate. It is worth noting that SGD is a simple stochastic algorithm, while more sophisticated optimizers are available in literature. Nevertheless, as it will be exposed in the case study section, it provides good results and then we assume there is no need to change it. At the end of each epoch, the loss is computed over the validation set in order to check the performance of the model over data which are not used for the training. Moreover, at each epoch the accuracy is computed on both training and validation sets, defined as the number of correct assignments of the model over the total size of the set. Plotting accuracy and loss versus the epochs results in the so-called learning curves. Visually, the learning curves could help in controlling the learning process: if their spread increases (i.e., training loss decreases more rapidly than validation loss), the model is occurring in overfitting, i.e., it is losing its ability to generalize and extract correct information from new data.

4.4.4 Test phase

Once the model is trained, it must be tested in order to evaluate its performance. The test set is used, which - by definition - has never been seen by the model in the

training phase. The following metrics are used in this phase:

1. Accuracy
2. Precision: for each class k, $p_k = \frac{\#\, data\, correctly\, assigned\, to\, k}{\#\, total\, data\, assigned\, to\, k}$. Precision gives an estimation of what is the proportion of the data assigned to class k that is effectively correct.
3. Recall: for each class k, $r_k = \frac{\#\, data\, correctly\, assigned\, to\, k}{\#total\, data\, belonging\, to\, k}$. Recall gives an estimation of the proportion of element belonging to class k that are found by the model.

After the testing phase, the model is ready to be used for the classification of new data.

Chapter 5

Hindu-Kush Karakoram/Himalaya seasonal cycle

The study area analyzed in this work is the Hindu-Kush Karakoram/Himalaya region (HKKH). This region could itself be divided into two distinct subregions, namely the Hindu-Kush Karakoram (hereinafter, HKK) and the Himalaya (Him) [Palazzi et al., 2013] (see figure 5.1). The interesting feature of this area is that the two subregions, even being space-contiguous, show considerably different seasonal precipitation patterns. As better detailed in the following paragraphs, the HKK region is characterized by a bimodal precipitation seasonal pattern, with a winter peak driven by Western Disturbances (WDs, see section 5.2) and a summer peak related to the Indian Summer Monsoon (section 5.1). The Him region, on the contrary, is only characterized by a summer peak, since the WDs contribution is confined further to the North-West. These differences in terms of seasonal cycles, widely documented in the literature, make the region a good case study for the validation of the proposed methodology for season identification.

In this chapter we will present a brief climatic characterization of the HKKH region, without the goal of being exhaustive. We will focus on the phenomenology of the precipitation seasonal features, rather than their physical drivers, trying to detect and validate the existing criteria found in literature for the time-space separation of the seasonal patterns. Attention is also paid to interannual variability, the possible role of atmospheric teleconnections, past and expected trends, in order to widen the number of criteria for the validation of the proposed methodology. Sections 5.1 and 5.2 deal with the Indian Summer Monsoon and the Western disturbances, respectively, while section 5.3 contains a focus on the HKK and Him regions. Section 5.4 contains the results of the proposed methodology for the region. Finally, in section 5.5 the results are briefly commented and discussed.

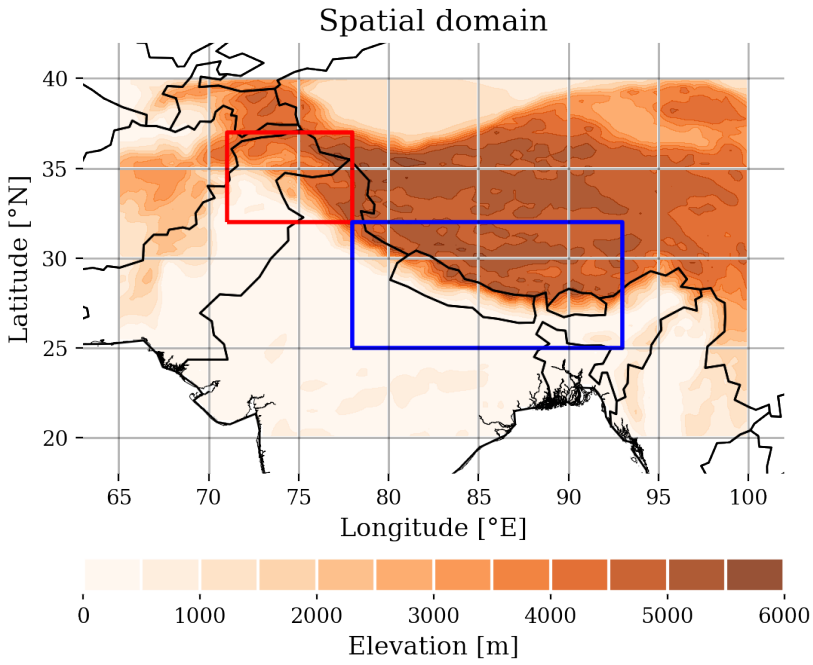

Figure 5.1: Spatial domain of the HKKH region: the red box represents the HKK box [Longitude 71–78 °E, Latitude 32–37 °N], the blue box represents the Him region [Longitude 78–93 °E, Latitude 25–32 °N]. Color shading shows the elevation data obtained from ERA5 orography.

5.1 Indian Summer Monsoon

5.1.1 Main features

The term "monsoon" is traditionally associated with the rainy period which accompanies a change in the seasonal prevailing wind in much of the tropics. In the regions prone to the monsoon, this circulation dominates the seasonal precipitation patterns, as the origin itself of the world, also suggests, which probably derives from the Arabic word *mausim* or the Malayan *monsin* which both mean season [Zhisheng et al., 2015]. For centuries it has been seen as a regional phenomenon similar to a giant land-sea breeze circulation [Gadgil, 2003, Zhisheng et al., 2015]. The more sophisticated concept of Global Monsoon (GM) emerged in the second half of the XX century, as global observational datasets became available. The GM could be interpreted as the first Empirical Orthogonal Function (EOF) of the annual anomaly of precipitation and circulation in the global tropics and subtropics, physically driven by the seasonal migration of the Intertropical Convergence Zone (ITCZ). Areas prone to GM are identified by the IPCC as those in which the annual precipitation range (i.e. the difference between the annual maximum and minimum precipitation) exceeds $2.5\frac{mm}{day}$, with no further requirements. However, this could lead to the inclusion of areas where the source of precipitation is not monsoonal. Therefore, a subsequent analysis was performed by the IPCC based on the published

literature [IPCC, 2021a] (figure 5.2).

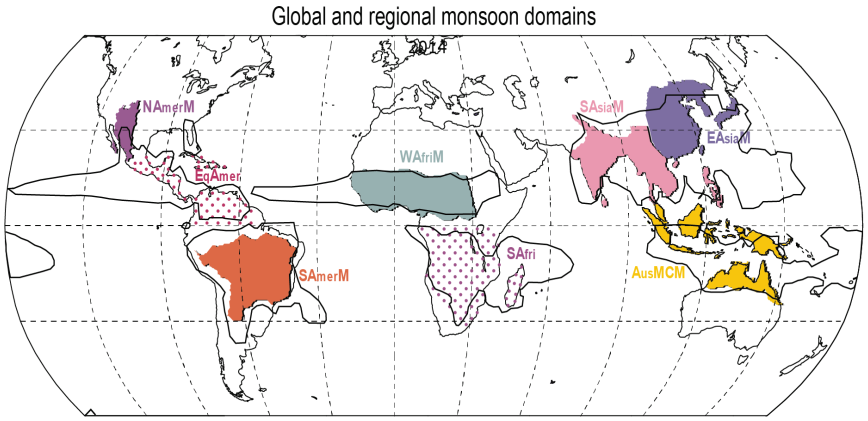

Figure 5.2: **Global and regional monsoon domains**: area interested by global monsoon (black line) and regional monsoon domains (colored areas). Regions that satisfy the GM criterium but are found to be dominated by a non-monsoonal dynamics are indicated with dots (source: IPCC, 2021: Annex V: Monsoons).

The South-Asian Monsoon (SAM) – the part of the monsoon system which mostly influences the HKKH region – can be regarded as part of the Asian Summer Monsoon (ASM). Since it covers wide geographical areas encompassing several countries, it has a unique impact on the economy of the region. In the Indian subcontinent, more than 60% of agriculture is rain fed and more than 70% of total rainfall occurs in the Monsoon season [Amrith, 2018]. Furthermore, the ASM provides precipitation to the southern slopes of Central and Eastern Himalayas. The spatial distribution of precipitation follows the orography of the region, with maxima located along the west coast of the Indian subcontinent (along the mountains called Western Ghats) and over the South-Eastern Himalayas [Gadgil, 2003]. The interannual variability of total precipitation shows consistent year-to-year fluctuations, while decadal variability presents alternate $\sim$ 30 years-long periods of precipitation above and below the average [Kripalani et al., 2003]. Until the end of the XX century a negative correlation between El Niño–Southern Oscillation (ENSO) phases (an irregular periodic positive anomaly in sea surface temperatures over the tropical eastern Pacific Ocean–one of the most important tropical teleconnections) and rainfall anomaly was observed, but in the last 20 years this relationship has shown a reversal suggesting the absence of a direct linkage [Gadgil, 2003, Dimri et al., 2016].

5.1.2 ASM Onset, progress, and withdrawal

Several studies, e.g. [Wu and Zhang, 1998, Liu et al., 2015], have found a triphasic space-time structure in the ASM onset process. The onset begins (1^{st} phase) in the

south-eastern part of the Bay of Bengal (known as the BOB monsoon), associated with an overturning of the meridional air temperature gradient [Mao and Wu, 2007] and the development of the so-called "monsoon onset vortex". This vortex, a low pressure system over the Central-East Arabian Seas, brings the monsoonal flow to the South-West Indian paninsula [Deepa and Oh, 2014]. This is followed (2^{nd} phase) by onset over the South China Sea, driven by atmospheric internal variability combined with the thermal and mechanical effects due to orography. The Indian Summer Monsoon (ISM) onset (3^{rd} phase) can be seen as the northward seasonal movement of the Intertropical Convergence Zone (ITCZ) [Gadgil, 2003], or alternatively, as the westward propagation of the BOB monsoon. In each onset phase, the importance of the thermal and mechanical effects of orography and particularly of the forcing associated with teh Tibetan Plateau has been highlighted several times [Liu et al., 2015]. It must be noticed that while the specific physical mechanisms are still open to debate, the triphasic structure is now commonly accepted by the scientific community. Since the physical characteristics of each component of the ASM is not relevant for the purpose of this work, we will hereafter simplify the discussion defining Monsoon the rainy season in India, in agreement with the India Meteorological Department (IMD).

After the onset, the Monsoon propagates north-westward and covers the entire Indian territory by middle of July [Pai and Rajeevan, 2009]. The withdrawal runs backwards the same trajectory, between the 15th September and the end of October. The Monsoon firstly hits the Indian South-Western state of Kerala. For this, the onset date is historically established looking at Kerala. In the last twenty years, the criteria for declaring the onset date have been updated several times, thanks to the availability of datasets with continuously increasing spatial and temporal resolution. Criteria used by the IMD and their historical evolution are briefly described here. It is worth noting that the IMD criteria are being widely used for the evaluation of new models.

For more than a century, the IMD has established the Monsoon onset date relying on seven rain gauge stations. The onset was declared on the second consecutive day after May 10^{th} in which measured rain exceeded $1\frac{mm}{day}$. This method was updated in 2006, when the current criteria were introduced [Pai and Rajeevan, 2009], in which the onset is declared over the Kerala state, after May 15^{th}:

- At least 60% of the 14 chosen stations rainfall values greater than or equal to 2.5 mm.

- Depth of westerlies should be maintained up to 600 hPa, in the geographical box extending from the equator to 10° N and from 55° E to 80° E. The zonal

wind speed over the area between 5 and 10^{o} N, and 70 and 80 °E should be of the order of 15–20 $Kts.$ at 925 $hPa.$ The data source can be either wind from analyses or satellite derived winds from the Regional Specialized Meteorological Centre for Tropical Cyclones over North Indian Ocean (RSMC)

- Indian National Satellite System (INSAT) derived Outgoing Longwave Radiation (OLR) value should be below 200 $\frac{W}{m^2}$ in the box confined by Latitude 5–10°N and Longitude 70–75°E.

After the onset over Kerala, the Monsoon advances northward across the subcontinent. Each region has its own onset date, called progress date, which is subsequential but not strictly correlated with the Kerala onset date (i.e., a delay in Kerala does not imply a delay in another country). In this section the so-called progress normal dates, which are the mean of the progress dates over a certain period, are presented.

Until 2020 the Monsoon progress normal dates were derived based on a network of 149 stations: the date of progress of monsoon over a station was taken as the middle date of the 5 days period showing the characteristic rise in the rainfall curve. The dates used by IMD were derived in the period 1901–1940. In 2020, a new method was introduced [Pai et al., 2020] based on a 1×1 gridded dataset (IMD-4) developed by the IMD and obtained by more than 2000 stations. This method was defined to obtain progress dates in agreement with the older method and was calibrated in the period 1961-2019 [Pai et al., 2020]. The Kerala onset date is the same as illustrated before. The other grid points are divided in 3 categories and for each category a specific method is defined. As a consequence, these methods are not physical but created ad-hoc.

In the literature, there are only a few studies about Monsoon withdrawal dates, especially compared with the studies focused on the onset and progress dates, and therefore the operational method has not been updated in 2020 [Pai et al., 2020]. The following criteria, adopted in 2006, are used, after September 1^{st}(IMD):

- End of rainfall activity over the area for five continuous days.
- Establishment of anticyclone in the lower troposphere (850 hPa and below).
- Considerable reduction in moisture content as inferred from satellite water vapor images and tephigrams.

5.1.3 Past and expected changes

In the last decades Indian Summer Monsoon has experienced a weakening in its circulation pattern and a decrease in its associated rainfall, which has been assessed

by many studies [Bingyi, 2005, Palazzi et al., 2013], probably caused by the warming of the Indian Ocean with a consequent decrease of sea-land temperature contrast. The role of this warming is still unclear and makes future projections uncertain: Global Circulation Models (GCMs) show a clear linkage between the increase in Sea Surface Temperature (SST) and the increase in monsoon rainfall, but the recent weakening of monsoon circulation seems to indicate that this temperature increase could result in a rainfall weakening, too [Roxy et al., 2015]. CMIP5 models indicate an increase in mean rainfall for the future, but show a significant inter-model spread in the representation of the seasonal cycle of rainfall patterns and only few models could reproduce it satisfactorily when compared to observations. On the other hand, the latest generation of CMIP6 models confirms the overall future trend found in CMIP5 and also shows a smaller internal spread, along with a better agreement with observations [Katzenberger et al., 2021].

5.2 Western Disturbances

5.2.1 Main features

The Indian summer Monsoon decreases while penetrating in the north-west of India and in northern Pakistan, and does not propagate far enough to reach internal Central Asian countries such as Afghanistan, Iran and Tajikistan [Seyed et al., 2006]. On the other hand, these areas are affected by recurrent events of winter precipitation (usually in the form of snow), which represent a precious water supply for the maintenance of glaciers and downstream for population [Seyed et al., 2006, Palazzi et al., 2013].

This precipitation pattern is caused by the so-called western weather patterns or Western Disturbances (WDs). The IMD defines them as the "extratropical storms that originate in the Mediterranean region which brings sudden winter rain to the north-western parts of the Indian subcontinent". These extratropical storms carry moisture in the upper layers of the atmosphere and then are pushed eastward by the westerly winds, until the interaction with the complex orography of the region leads to precipitation. Although the mechanism of this simplified model seems straightforward, the phenomenon has not been fully understood yet [Dimri et al., 2016]. It has been found that WDs have an important impact on the Summer Monsoon, since they induce a local change in albedo through snow accumulation on the mountain ranges, and albedo has an important role in the development of the Monsoon. Nevertheless, a full understanding of the phenomenon is far to be reached [Dimri et al., 2016].

WD precipitation shows a large interannual variability, and has been found to

weaken in presence of a positive anomaly of SST in the Arabian Sea. Furthermore, WD precipitation shows a positive anomaly associated with a positive NAO and a warm ENSO, and a negative anomaly with negative NAO and cold ENSO [Dimri et al., 2016]. Contrary to the case of the monsoon, WDs do not have a large literature background concerning their timing (i.e. onset and decay) and evolution. For this reason, we will keep as reference for their typical period the standard five months from December to April (DJFMA) as in e.g. [Palazzi et al., 2013].

5.2.2 Past and expected changes

The area interested by WDs (Western Himalayas) has already experienced a significant trend in increasing temperatures in the last decades. This especially affected the mountain areas, making the region a case study of the so-called "elevation-dependent warming", i.e. the emerging evidence that mountain environments around the world are experiencing a more rapid change in temperature. In the period 1961–2006, the observed warming was of 2-2.5 °C above 5000 m and only 0.5 °C at sea level [Xu and Rutledge, 2019]. On the other hand, summer cooling has been reported for the period 1961–2015 [Krishnan, 2019], with an associated thickening of the local glaciers. Historical precipitation trends have not been defined, especially due to the lack of stations in the region [Palazzi et al., 2013].

In the future, temperature is expected to increase with a high level of confidence, with the possibility of exceeding an increase of 5°C by the end of the century in the SSP5-8.5 high emission scenario [IPCC, 2021b]. Projections on precipitation are more uncertain. CMIP5 models exhibited some spread in the representation of the precipitation seasonal cycle in the region [Palazzi et al., 2015], and also dynamically-downscaled datasets showed a similar behavior [IPCC, 2021b]. However, the new generation of CMIP6 GCMs shows an increment in winter precipitation, assessed with medium level of confidence in the IPCC AR6 [IPCC, 2021b].

5.3 Seasonal cycle in the HKKH

In this section the seasonal precipitation pattern in the study region is evaluated. We use the spatial division performed by [Palazzi et al., 2013, Palazzi et al., 2015] to define climatically-coherent subregions of the entire HKKH: the Hindu-Kush Karakoram (HKK) and Himalaya (Him) regions (as shown by Figure 5.1). These boxes have been created since the spatial features of the Monsoon and the WDs prevent us from treating the HKKH as a single region, as it is exposed to different circulation patterns affecting precipitation seasonality. In fact, the Himalayan region is

dominated by Monsoon-controlled dynamics, while in the HKK, precipitation also occurs during Winter, due to the WDs. In section 5.3.1 we evaluate the spatial behavior of the seasonal precipitation pattern in the HKKH region in the ERA5 Reanalysis and in the EC-Earth3 Earth System Model, i.e. in the datasets that will be used for the evaluation of our methodology for seasons identification. Note that for EC-Earth3 we have three ensemble members (see chapter Data). In this part we will use the mean of these members, usually known as ensemble mean. Section 5.3.2 contains a literature review of the seasons onset and withdrawal, that from now on we will also call **seasonal breakpoints**. The aim of this section is to establish the reference seasonal breakpoints that will be used for the validation of the model. Finally, in section 5.3.3 the future trends in precipitation seasonal patterns will be assessed, using as seasonal breakpoints the ones identified in section 5.3.2. The results obtained in this part will be compared with the future trends obtained with the seasonal breakpoints that will result from our algorithm.

As highlighted, the most interesting seasonal pattern in the region is the one concerning precipitation. Thus, the main focus of our analysis will be on this variable. The analysis presented in this section is performed on precipitation, and also the result of our methodology for the division in seasons will be discussed focusing on precipitation. Nevertheless, we could assume that precipitation alone is not enough for the evaluation of seasons. For example, Monsoon onset over Kerala is determined looking also at winds and OLR, as detailed before. On the other hand, using a set of ad-hoc chosen variables for each case study would be in contrast with our purpose of generality. Therefore we decided to use for our model also the surface air temperature. This choice is driven by the fact that total precipitation and surface air temperature are the most used variables for climatic characterization. Spatially, surface air temperature in HKKH presents a gradient that follows the elevation. During the year, the maxima are located in Summer and the minima in Winter. In this work we decided to omit an in-depth analysis of temperature, which could be explored in a followup work.

Note on nomenclature: in the following sections, we will introduce three sets of seasonal breakpoints, i. e., dates which mark the transition between the seasons. The first is obtained from a review performed on the literature, the second will be the result of our clustering algorithm. For clarity, from now on we will call **reference breakpoints** the first set, and **algorithm breakpoint** the second one. As a consequence of the definition we introduced, these breakpoints are static, which means that they do not change from one year to another. Consequently, a third set is introduced, and is used for the future evolution of meteorological seasons, obtained with classification. These ones will be defined as of **dynamical breakpoints**, since

these breakpoints can change through years.

5.3.1 Evaluation of the HKK and Him precipitation

As seen in the previous section, rainfall in the North and Northwest areas of India are affected by the northwestward propagation of the Monsoon and by the eastward propagation of the Westerly Disturbances, which both weaken in their paths. This results in a division in two areas: one where the Winter peak is more prominent than the Summer one, and the other one where the opposite occurs. An evaluation of this division in the ERA5 dataset shows that the HKK box includes the area dominated by the WDs characterized by a significant precipitation amount, both in Winter and Summer (Figure 5.3 A-B). On the other hand, the Him box includes the areas where the Summer Monsoon is the dominant feature, and excludes the Eastern sector over Bangladesh, where there is also a significant peak during winter. An effective method to extract the different role of the monsoon and WDs is to compare the intensity of the winter and summer peak (Figure 5.3 C). We can note that the Him box includes the area dominated by the Summer peak while the HKK box is mostly dominated by Winter peak, except for the South-Western part. In the EC-Earth3 climate model, the situation is slightly different. Here the difference of the peaks is evaluated using the whole available period (1850–2100) (Figure 5.4). The Him box is dominated by the Summer peak with no significant variations over time. The HKK box is dominated by Winter peak, but its intensity is lower than in ERA5. Nevertheless, also for HKK no significant variations over time emerge.

Now we evaluate the mean seasonal cycle in the two boxes for both datasets. For ERA5, as expected, the mean precipitation seasonal pattern is bimodal with two peaks in the HKK box (Figure 5.5 A), and has only one peak in the Him region (Figure 5.5 B). For EC-Earth3, it is possible to note a delay of about one month in both the summer and winter peak, compared to the seasonal cycle in ERA5, in the HKK region for the historical period (Figure 5.6 A). Furthermore, a small peak can be observed in November, probably due to a specific bias in the EC-Earth3 seasonal precipitations cycle, which waa also already assessed by [Palazzi et al., 2015]. In EC-Earth3 future projections (SSP5-8.5 scenario) for the HKK region (Figure 5.7 A), the winter peak is replaced by a plateau that extends from March to May, while the summer peak anticipates compared to the historical period. As for the Him region, the Summer peak in both historical and future simulations of EC-Earth3 is slightly delayed compared to ERA5 (Figures 5.6 B and 5.7 B). Overall, EC-Earth3 shows a dry bias of about 2 mm/day compared to ERA5.

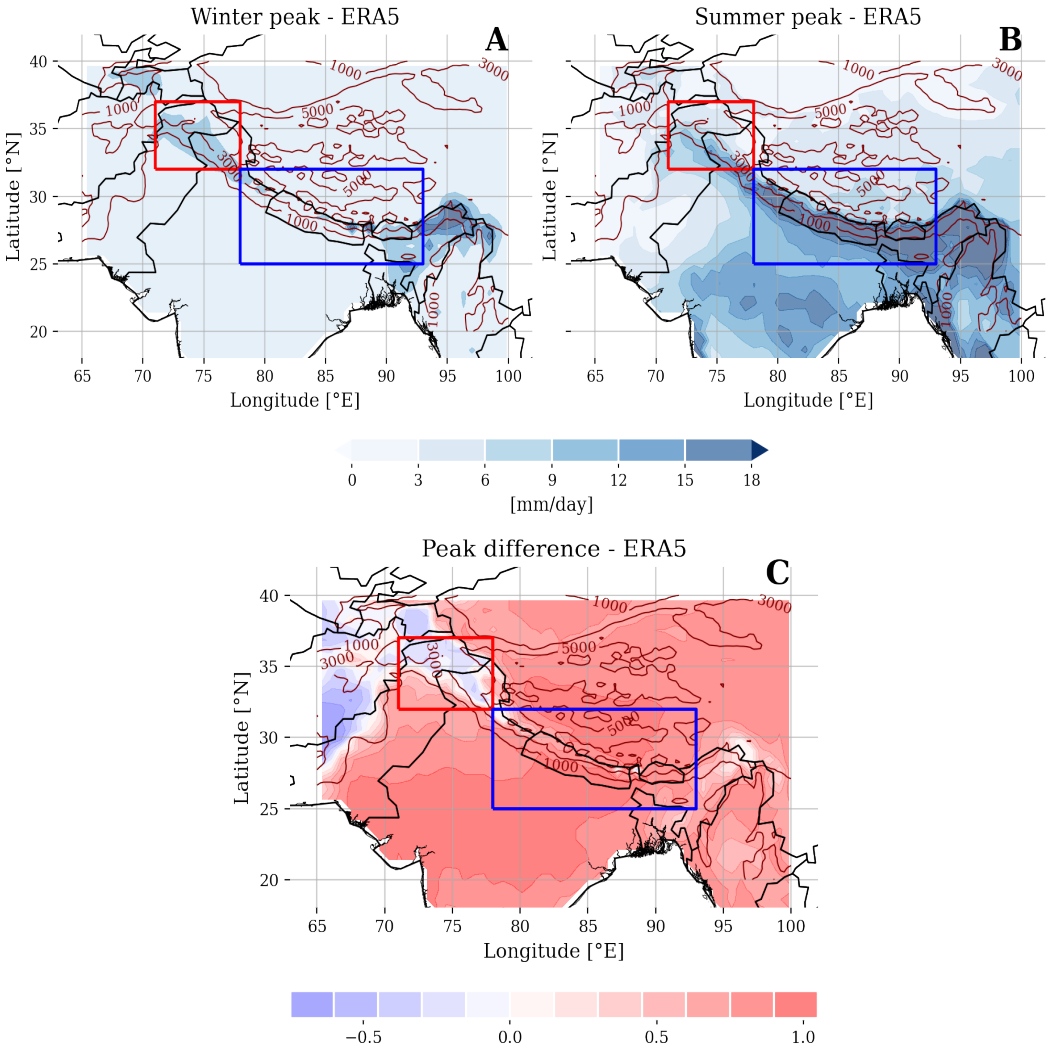

Figure 5.3: **Summer and Winter peak in HKKH in ERA5** (1979-2020): maximum of total precipitation seasonal cycle in winter months, i.e. NDJFMA (A), maximum of total precipitation seasonal cycle in summer months, i.e. MJJASO (B), difference between the maximum in NDJFMA and MJJASO (C). In the peack difference the seasonal cycle of each grid point has been previously normalized with min-max normalization, in order to compare the intensity of peaks. Thus blue areas are dominated by winter peak, and red areas by summer peak. For each graph ERA5 in the period 1979-2020 has been used. Red contours, shown in all panels, represent the orography obtained by ERA5, with [m] as unit for the inline values.

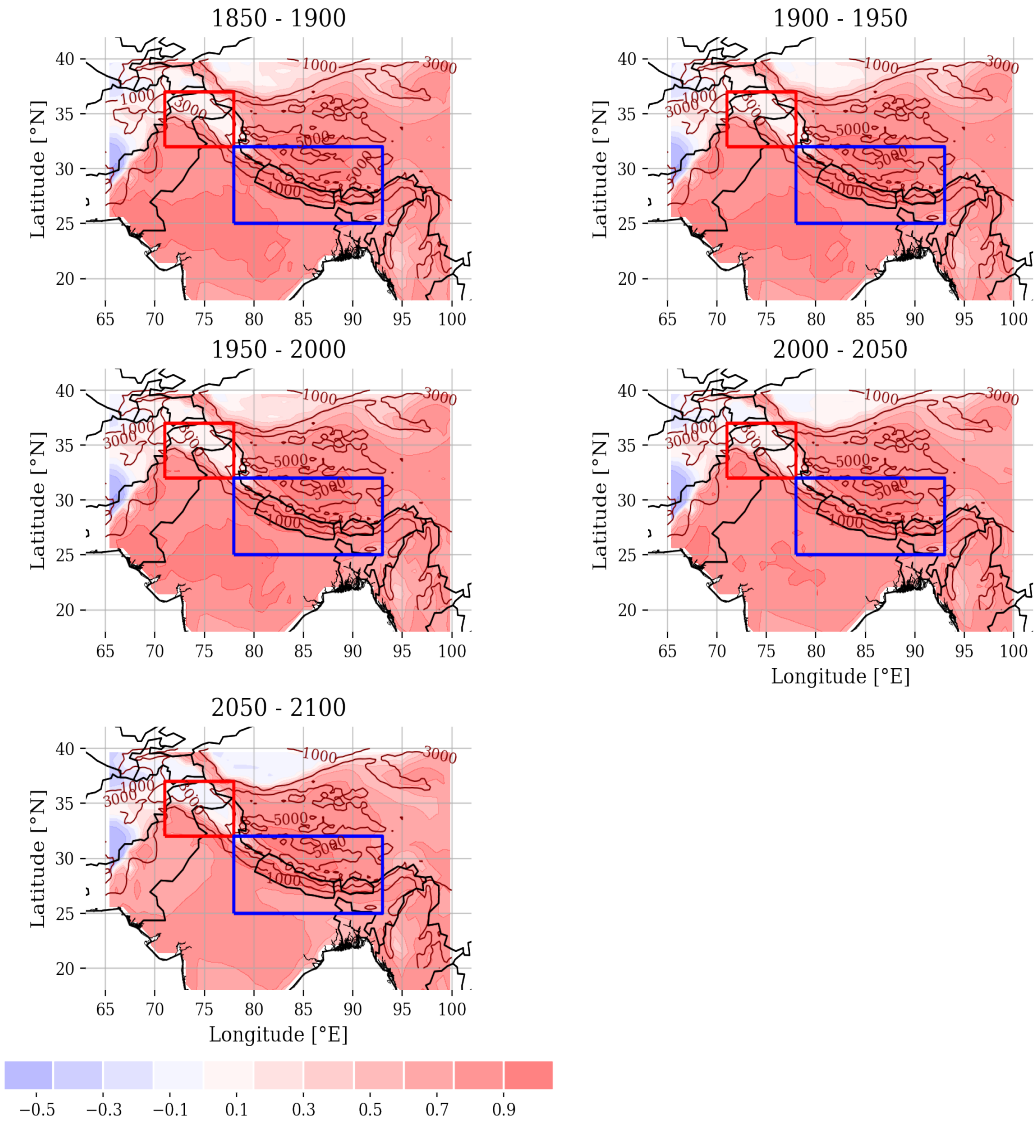

Figure 5.4: **Summer and Winter peak in HKKH in EC-Earth3** (1850-2100): same as Figure 5.4 but for EC-Earth3 climate model on different time windows.

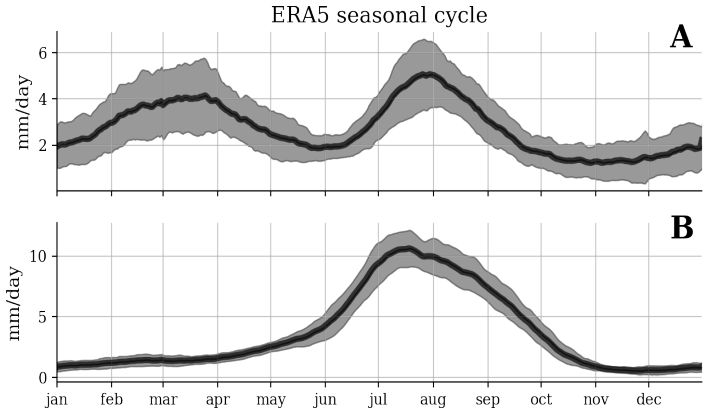

Figure 5.5: **Seasonal precipitation cycles in HKK (A) and Him (B) boxes in ERA5 (1979-2020)**. The seasonal cycle is computed averaging precipitation for each ordinal day of the year. Solid lines are the spatial mean, while shadowed areas are the spatial standard deviation.

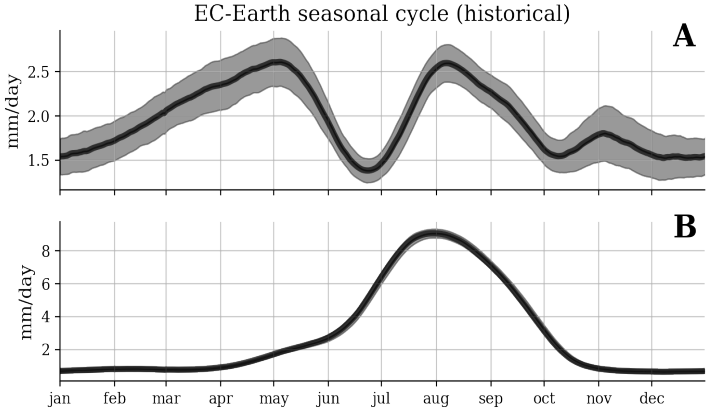

Figure 5.6: **Seasonal precipitation cycles in HKK (A) and Him (B) boxes in EC-Earth3 historical (1850-2014)**. Same as Figure 5.5 but for EC-Earth3 historical.

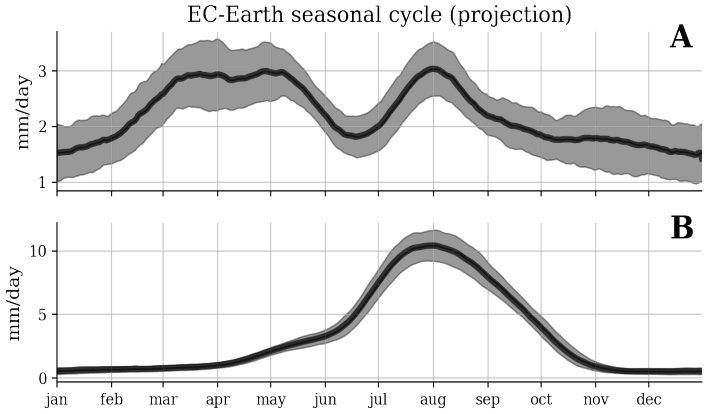

Figure 5.7: **Seasonal precipitation cycles in HKK (A) and Him (B) boxes in EC-Earth3 SSP5-8.5 scenario (2015-2100)**. Same as Figure 5.5 but for EC-Earth3 SSP5-8.5.

5.3.2 Breakpoint dates review

For the reference breakpoints, the literature suggests DJFMA as the winter precipitation season and JJAS as the summer monsoon season. e.g. [Palazzi et al., 2013]. The two intermediate dry seasons turn out to be only May for Spring and October and November for Autumn. A finer approach to timing could be performed considering the normal onset and withdrawal dates proposed by the IMD (Figure 5.8). The onset date within the Him boundaries is June 8^{th} $\pm$ 7 days, the withdrawal date is October 5^{th} $\pm$ 6 days for a total length of 109 $\pm$ 11 days. These dates will be used as reference for the Him box, since a two-season model seems to be the best approach for this region. For the HKK region the breakpoints suggested by [Palazzi et al., 2013] will be used for the transition from Autumn to Winter and from Winter to spring, while we will use the IMD dates for on the onset and withdrawal date of the Monsoon. IMD dates in HKK boxes are June 25^{th} $\pm$ 3 days for the onset and October 2^{nd} $\pm$ 1 day for the withdrawal, with a total length of 89 $\pm$ 4 days. Table 5.1 summarizes the reference breakpoints we will use for the model validation.

	Winter	**Spring**	**Summer**	**Autumn**
HKK	1 Dec - 30 Apr	1 May - 25 Jun	25 Jun - 2 Oct	3 Oct - 31 Nov
	Dry		**Monsoon**	
Him	6 Oct - 17 Jun		18 jun - 5 Oct	

Table 5.1: **Reference breakpoints** in HKK and Him boxes based on literature review. The seasons names have been chosen arbitrarily, and don't necessarily have references to the seasons at mid-latitudes.

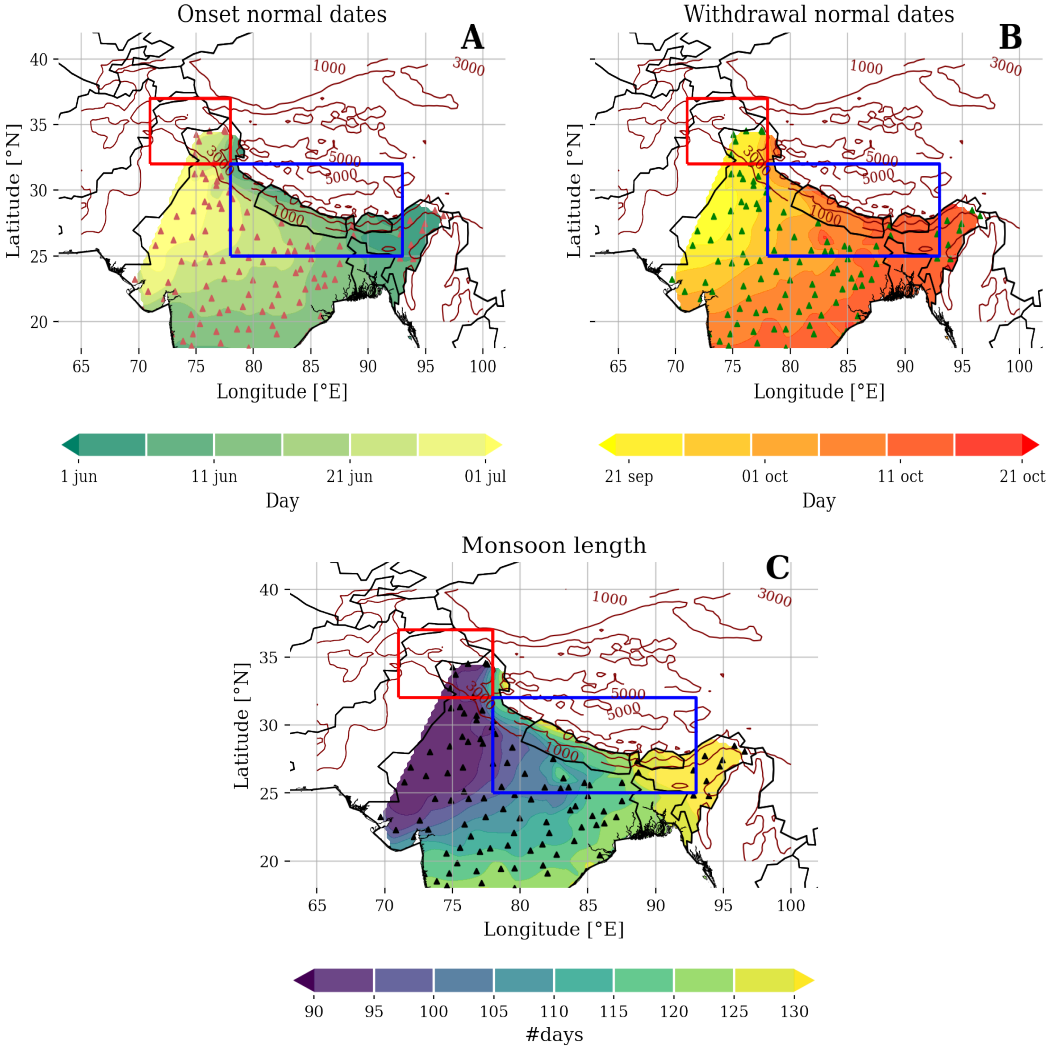

Figure 5.8: **Monsoon onset and withdrawal**: summer monsoon onset normal date (A), summer monsoon withdrawal normal date (B) and resulting length of the monsoon season (C). Figures are obtained by bilinear interpolation of multiple 'single station' values provided by IMD. These points are represented by triangles. Red contours are orography obtained by ERA5, with [m] as unit for the inline values.

5.3.3 Future trends

Now the rainfall future trends (2020-2100) for multi-members mean EC-Earth3 model for the case study regions making use of the reference breakpoints are evaluated. Mean daily precipitation and seasonal accumulated precipitation are chosen as reference metrics for each season since both of them are sensitive to the seasonal boundaries. In fact, given the sinusoidal shape of the seasonal precipitation pattern, changing the boundaries modifies the mean and cumulative values. For this reason, they are suitable for the comparison between trends obtained with reference breakpoints and the ones obtained with algorithm breakpoints. Here we will use a simple linear regression for trends evaluation. Please bear in mind that there is a wide spectrum of available methods for the assessment of trends in climatology, such as nonlinear methods, or nonparametric methods, and the most proper one should be chosen based on the specific application.

The choice of a linear regression allows us to make a simple consideration about the relationship between the trend of mean values and cumulative values. If we assume that the number of days in a season is fixed, there is only a multiplicative difference in trends of mean and cumulative values: being m_t the time series of mean values, c_t the time series of cumulative values, L the linear operator representing the linear trend, and n_t the number of days in a season, the trends for m_t and c_t will be respectively:

$$T_m = L[m_t]$$
$$T_c = L[c_t]$$

But being $m_t = c_t * n_t$:

$$T_c = L[m_t * n_t]$$

If n_t is constant in time, $n_t = n$, for linearity we can write:

$$T_c = L[m_t] * n = T_m * n$$

The last identity is not true if n_t can change through years, meaning that in this case there will not be a linear dependency between the two metrics. The reference breakpoints, as the algorithm breakpoints, are time invariant, i. e. they do not change throughout the years. So, with these breakpoints the number of days in a season is fixed and there is a linear relation between mean values and cumulative value trends. The dynamical breakpoints could change through the years. So, the number of days in a season is not fixed and this linear relation is no longer guaranteed. Now the results obtained with reference breakpoints are presented. Note that being mean values trends and cumulative values trends linearly related,

presenting both of them would be redundant. For this reason, now only the mean values trends are presented.

With the reference breakpoints, an overall precipitation increase in HKK is expected in the period 2020-2100 under the SSP5-8.5 scenario, except for Autumn (Figure 5.9 D). This increase is particularly pronounced in the South-East area in Summer and North-West in Winter (Figure 5.10 A, C), suggesting an increase of precipitation associated respectively to Monsoon and WDs. Also in Him box positive trends are expected, especially in mountain areas during Monsoon season (Figures 5.11 and 5.12). A spot of positive significant trends in South-East area in the Dry season (Figure 5.12 A) suggests caution, since this is the area that is firstly hit by monsoon and last left by its withdrawal. This trend may be therefore caused by a stretching of the monsoon season, and will be verified at the light of the results obtained with dynamical breakpoints.

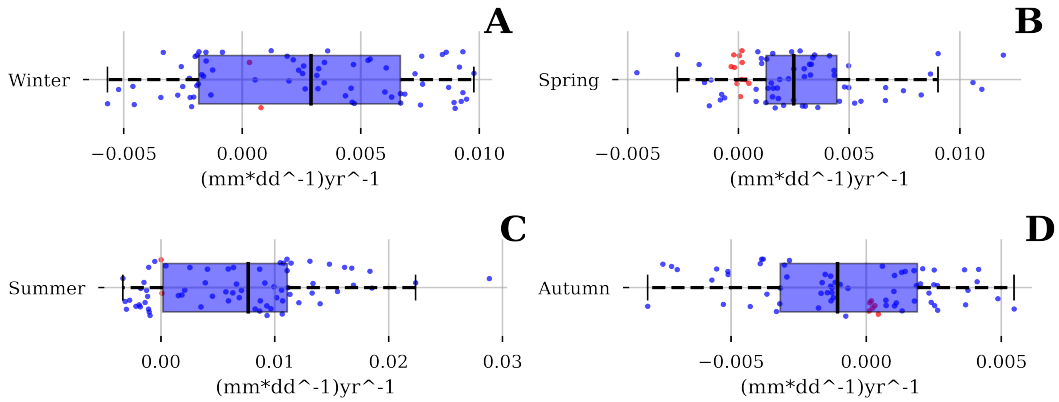

Figure 5.9: **HKK mean seasonal precipitation future trend boxplot (reference breakpoints)**: linear trends are computed with a Mann-Kendall test for monotonic trend over each grid point. Trends are computed for each season defined with reference breakpoints separately, using data contained in EC-Earth3 dataset for future projection under SSP5-8.5 scenario. Each scatter point represents the trend of a grid point. Red markers indicate points whose p-value exceeds the threshold of 0.05, and are therefore considered non significant. The boxplot only represents significant values.

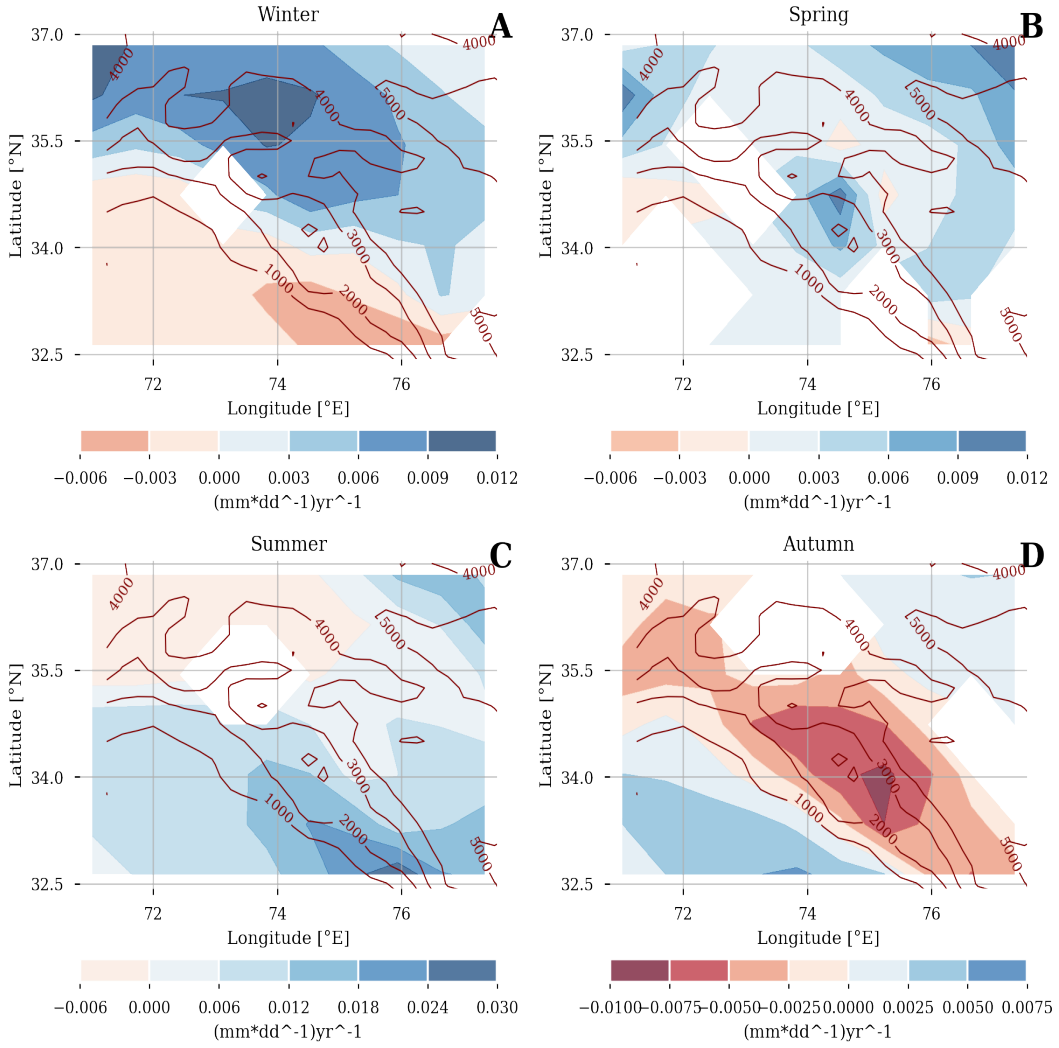

Figure 5.10: **HKK seasonal mean precipitation future trend (reference breakpoints)**: the method is the same of figure 5.9. Here non-significant values are shown in white. Red contours, shown in all panels, represent the orography obtained by ERA5, with [m] as unit for the inline values.

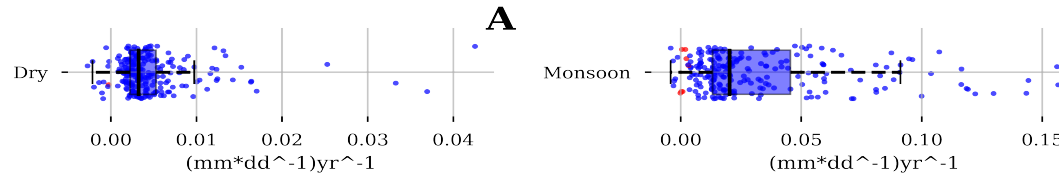

Figure 5.11: **Him mean seasonal precipitation future trend boxplot (reference breakpoints)**: same as figure 5.9.

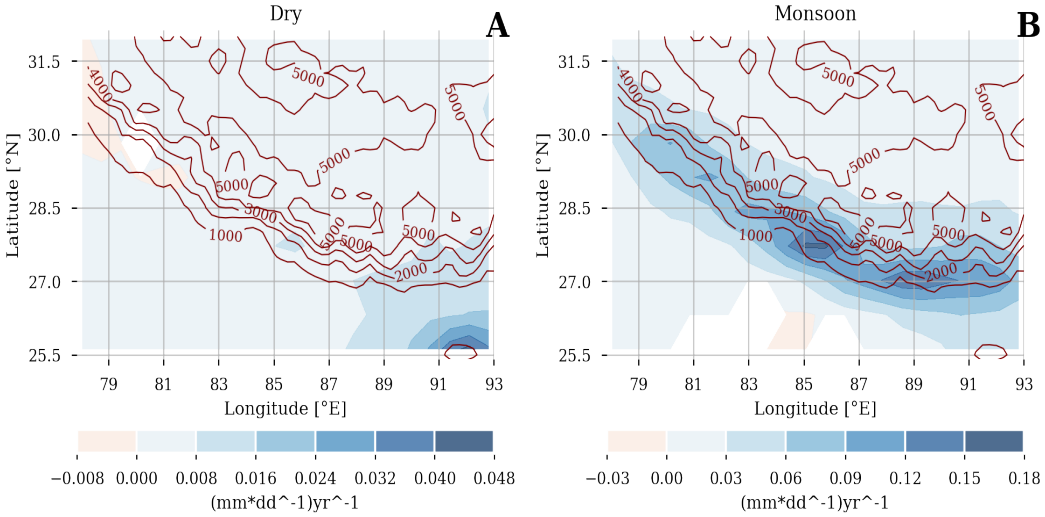

Figure 5.12: **Him seasonal mean precipitation future trend (reference breakpoints)**: same as figure 5.25.

5.4 Results of the model

This section contains the application of the algorithm designed in this thesis (Chapter 3) to the HKK and Him regions. As detailed before, for the identification of seasons both surface air temperature and total precipitation are used, but the results are exposed focusing on precipitation. In the first part, the Radially Constrained Clustering algorithm is used on the ERA5 dataset over the period 1979-2020. This will allow for the identification of the algorithm breakpoints, i.e. the dates which mark the transition between seasons defined in a data-driven way. These breakpoints will be compared with reference breakpoints to assess the results of the algorithm. The meteorological seasons obtained in this part will be used as base ground for the assessment of the seasons in the EC-Earth3 future climate simulations. This will be done in the second part of this section, where the classification of EC-Earth3 data into the seasons is performed making use of the SoftMax perceptron. The classification will be performed for each ensemble member both in historical period and future projection with SSP5-8.5 scenario. This will lead to the identification of the dynamical breakpoints. There are three scientific questions we will try to answer in this part:

1. How much the EC-Earth3 representation of the seasons could be considered reliable, compared with ERA5 one. This will be addressed by comparing the algorithm breakpoints with dynamical breakpoints in historical period.

2. How meteorological seasons have changed in the historical period, and how they will change in the future, according to EC-Earth3. This will be addressed evaluating the time evolution of dynamical breakpoints.

3. How an evolving recognition of meteorological seasons influences the changes in future season-dependent precipitation. and the related trends in mean and cumulated seasonal values. This will be addressed computing the trends with dynamical breakpoints and comparing them with the ones obtained with reference breakpoints.

In section 5.4.1 the optimal number of seasons is estimated using a combination of the elbow method and silhouette score, to validate the goodness of the number of seasons found in literature. In section 5.4.2 the clustering is performed. Both these operations are carried out on ERA5 dataset over the period 1979-2020. In section 5.4.4, EC-Earth3 data are classified. Finally, in section 5.5, the results are briefly discussed.

5.4.1 Number of seasons

The ideal number of seasons is evaluated using the ERA5 dataset. Radially Constrained Clustering (RCC) is performed for a number of clusters in range [1; 10] and for each result the total Within Sum of Squares (WSS) and the silhouette coefficient are computed. WSSs are reported in the elbow graph and tend to decrease when increasing the number of clusters. Thus, the optimal number is usually chosen taking the point after which the gain decreases, i.e. the elbow of the graph. Silhouette score compares the intra-clusters distance with the inter-clusters distance and has 1 as its optimal value.

According to the silhouette score, for the HKK box an optimal number of clusters seems to be $N = [7, 9]$, since they are the values that achieve the highest scores (Figure 5.13 B) meaning that clusters are well differentiated. On the other hand, this will result in a high number of seasons which would be of a length of about less than two months. A 4 seasons analysis, as found in literature, seems to be a good approach in elbow graph (Figure 5.13 A), but achieves a bad score in silhouette (Figure 5.13 B), meaning that the clusters are not well differentiated. A good compromise between the two metrics seems to be $N = 6$: elbow graph shows a decrease in gain and silhouette achieves a relatively good score. Nevertheless, the algorithm breakpoints achieve better scores in both elbow graph and silhouette compared to reference breakpoints (Figure 5.13 A B). In the following of this work, we will continue to use N=4 in order to get results comparable to literature. A future development of this work could be done varying the number of seasons, using a value that is more performing in the metrics.

In the Him region both metrics suggest that a 2 seasons clustering is the best approach: silhouette achieves the best score and the elbow in elbow graph is clearly

distinguishable, meaning that the clusters are well defined and differentiated. Being two also the reference breakpoints, we will continue our analysis with this number of seasons.

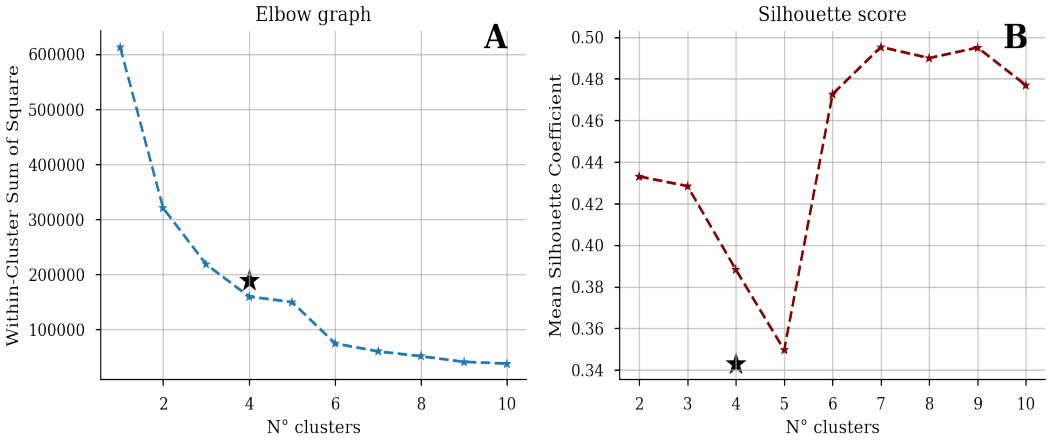

Figure 5.13: **HKK number of seasons metrics**: dashed lines are metrics computed with RCC algorithm, black stars are the metrics computed on the clusters obtained with the reference breakpoints

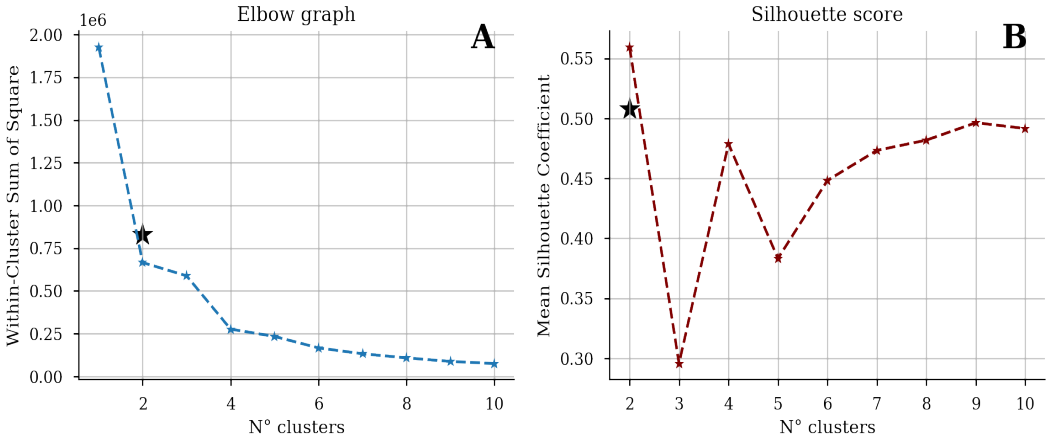

Figure 5.14: **Him number of seasons metrics**: dashed lines are metrics computed with RCC algorithm, black stars are the metrics computed on the clusters obtained with the reference breakpoints

5.4.2 Clustering results

In this section the RCC algorithm is used on ERA5 dataset to compute the algorithm breakpoints in order to obtain the data-driven seasons. This operation is performed with 4 clusters for HKK and 2 clusters in Him, as documented in the previous section.

Clustering performed over the HKK region shows a good agreement with reference breakpoints (Figure 5.15). The succession of seasons is respected, and the length is almost everywhere similar to the one described by reference breakpoints. The result on the transition from Spring to Summer is particularly remarkable: this breakpoint is assessed by many studies in literature and could be reproduced with great precision by the model. The discrepancy with Winter withdrawal can be traced back to the fact that this reference breakpoint does not recognize units smaller than 1 month. On the other hand, the starting Winter breakpoint shows an advance of 1 month from the reference one.

In the Him box the cluster performed with 2 seasons shows a good agreement with the reference breakpoint in the transition from Monsoon to Dry, but has a shift of about one month in the transition from Dry to Monsoon (Figure 5.16). In Table 5.2 the reference and algorithm breakpoints are reported.

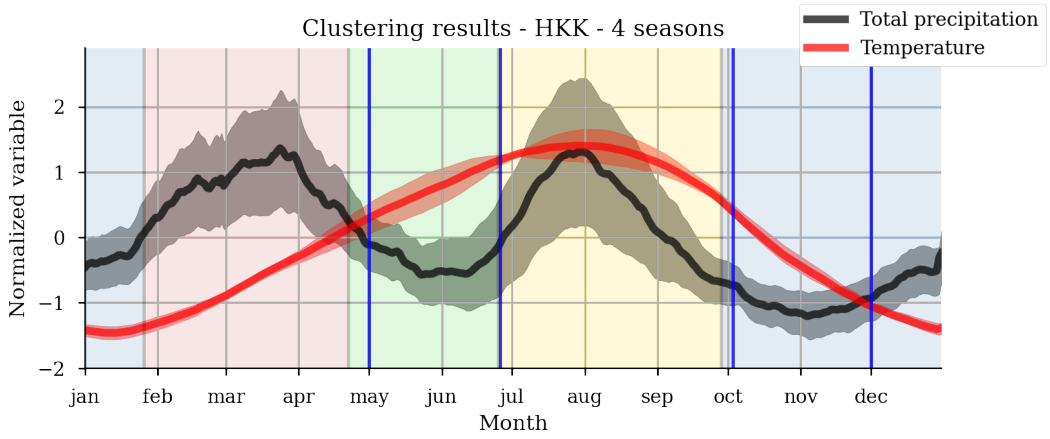

Figure 5.15: **Clustering results in HKK**: solid lines are values averaged on the whole region, shadowed areas are spatial standard deviations. Clustering results are reported as background colors, while blue lines are reference breakpoints.

		Winter	Spring	Summer	Autumn
HKK	Algorithm	24 Jan - 20 Apr	21 Apr - 25 Jun	25 Jun - 27 Sep	28 Sep - 23 Jan
	Reference	1 Dec - 30 Apr	1 May - 25 Jun	25 Jun - 2 Oct	3 Oct - 31 Nov
		Dry		**Monsoon**	
Him	Algorithm	1 Oct - 9 May		10 May - 30 Sep	
	Reference	6 Oct - 17 Jun		18 jun - 5 Oct	

Table 5.2: **Reference and algorithm breakpoints** in HKK and Him boxes based on literature review and on the results of RCC algorithm. The seasons names for RCC results have been assigned arbitrarly.

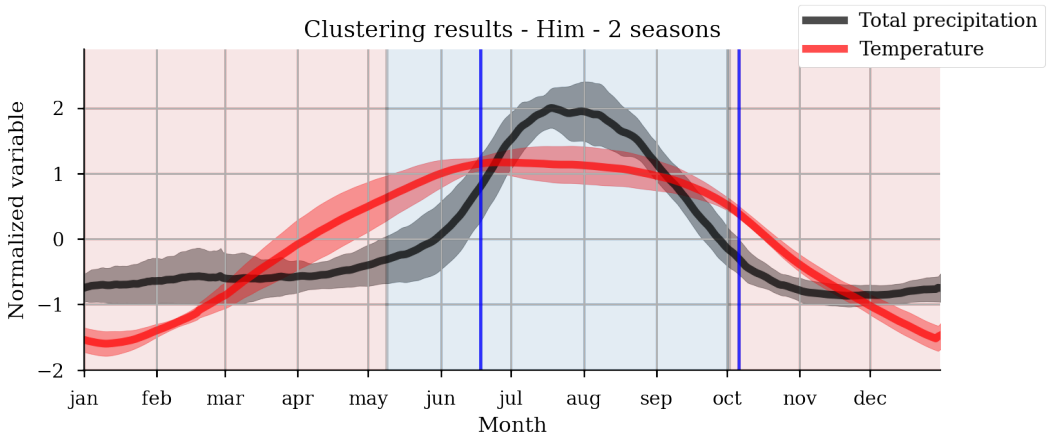

Figure 5.16: **Clustering results in Him**: solid lines are values averaged on the whole region, shadowed areas are spatial standard deviations. Clustering results are reported as background colors, while blue lines are reference breakpoints.

5.4.3 Training of the SoftMax perceptron

In order to proceed with the classification of the EC-Earth3 data, the SoftMax perceptron must be trained and tested with the labeled dataset obtained in the previous section.

For the HKK box, the SoftMax perceptron is setted with 140 input units (2 variables * 10 lon points * 7 lat points) and 4 units in the output layer, corresponding respectively to Winter, Spring, Summer and Autumn seasons. The quality of the perceptron learning will be assessed making use of the metrics described in Chapter Methods.Looking at the learning curves, we can assume that 50 epochs are enough for the training. Indeed, the accuracy reached stability (Figure 5.17 B) even after 30 epochs. The loss seems to be able to decrease further (Figure 5.17 A), but since there is no improvement in accuracy we assume that it is not necessary to increase the number of epochs. From the test phase we can state that the model is able to learn the relationships between data and seasons, achieving good scores with fresh new data (Figure 5.18). In fact, precision and recall are > 0.99 for all the seasons.

For the Him box, the SoftMax perceptron is setted with 440 input units (2 variables * 20 lon points * 11 lat points) and 2 units in the output layer, corresponding respectively to Dry season and Monsoon season. The learning curves suggest that 40 epochs are enough for the training. Also in this case, despite loss seems to decrease further (Figure 5.19 A), accuracy has reached stability (Figure 5.19 B). Again, the confusion matrix states that the model has a good performance (Figure 5.20), with precision and recall > 0.99.

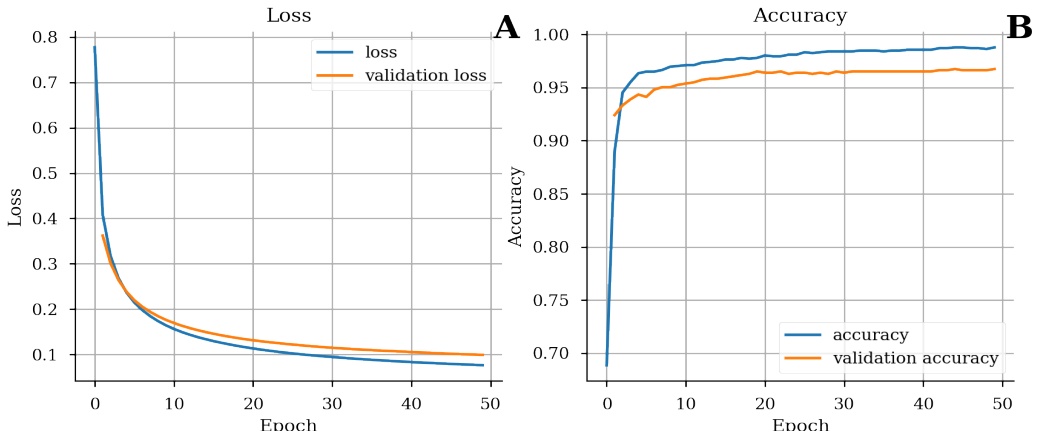

Figure 5.17: **Learning curves for HKK**: loss on training and validation sets (A), accuracy on training and validation sets (B).

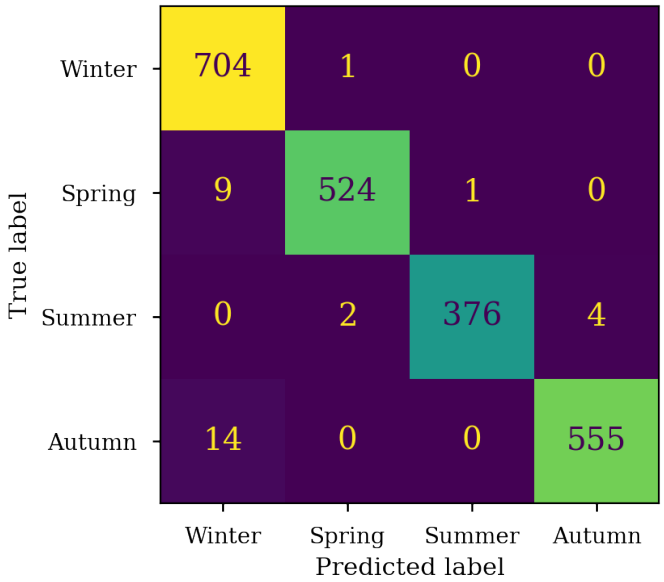

Figure 5.18: **Confusion matrix for HKK**: number of items belonging to each class versus number of items classified in each season. The elements on the diagional are the items correctly classified.

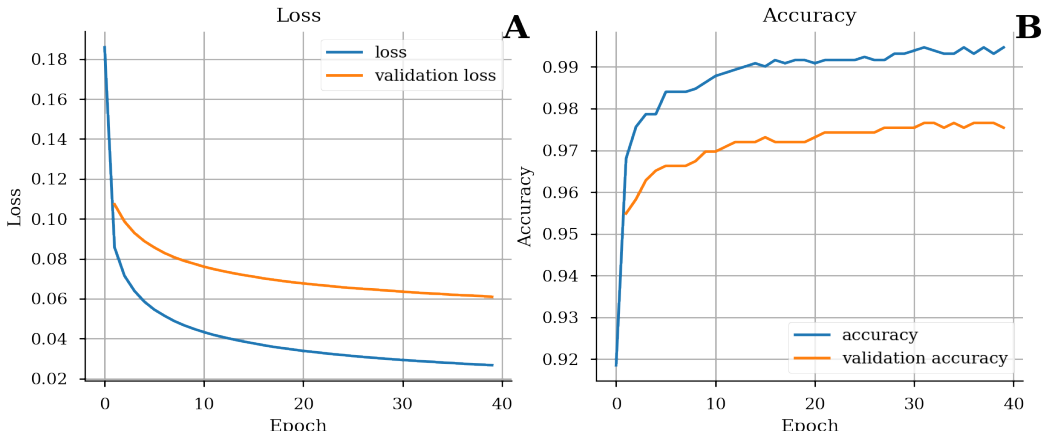

Figure 5.19: **Learning curves for Him**: loss on training and validation sets (A), accuracy on training and validation sets (B).

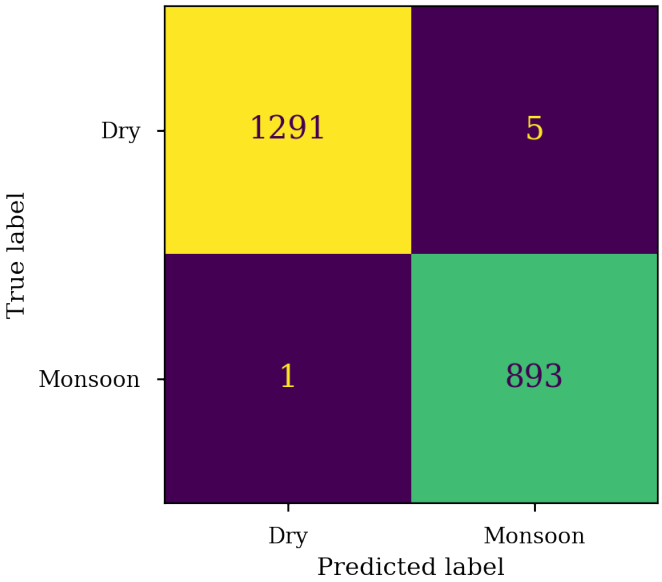

Figure 5.20: **Confusion matrix for Him**: number of items belonging to each class versus number of items classified in each season. The elements on the diagional are the items correctly classified.

5.4.4 Results of classification on climate projections

Now the SoftMax perceptron is used to classify the data in EC-Earth3. As detailed in Methods chapter, for both HKK and Him boxes, the classification performed with the SoftMax perceptron assigns to each day in EC-Earth3 the probability that this day belongs to each season separately. As explained in the Dataset chapter, we have three ensemble members of EC-Earth3 (r1i1p1f1, r13i1p1f1, r15i1p1f1). Each of them is composed of the historical period (1850-2014) and the future projection under the SSP5-8.5 scenario (2015-2100). The classification is performed for each day of each ensemble member, in order to obtain three possible realizations of the meteorological seasons, which spans from 1850 to 2100. This range is reduced due to the 30 years moving average performed on seasonal cycle (see Dataset chapter), so that the actual results span from 1865 to 2085.

An overall result is obtained by averaging this probability through ensemble realization. That is to say, for each day the probability that it belongs to each season is average through ensemble members. Thus we can assign each day to the season that achieves the highest probability (the so-called arg max mathematical function). The arg max is one of the possible interpretations of this probabilistic output. The result is a single breakdown of the days in the meteorological season. A future development of this work could try to explore more powerful methods to extract information from this type of result.

This operation is performed for both HKK and Him boxes. The first result is that for both the regions the succession of seasons is respected quite evenly (Figures 5.21 and 5.22). This is non-obvious since the SoftMax perceptron does not have information about the time location of each day on the calendar year. This fact means that the SoftMax perceptron is able to extract relevant features from the ERA5 dataset, and that these features are correctly recognized in EC-Earth3.

The Autumn and Winter seasons in HKK are an exception to this result (Figure 5.21). They are quite fragmented and a relevant spot of Winter day could be found in the middle of the Autumn, especially in the period 1900-2020. This is probably due to the third peak that was observed in Figure (5.6), and which was stated to be an error of EC-Earth3. In this sense, this result could help in locating inconsistency in EC-Earth3 seasonal cycle representation. The fragmentation of Winter and Autumn boundaries in HKK could be due to the fact that these seasons are not well differentiated. In fact, we stated in section 5.4.1 that four seasons is not the optimal choice for the HKK. A further development of this work could investigate if changing the number of seasons produces more stable results in this sense.

In Him box the seasons are contiguous and the boundaries are not fragmented (Figure 5.22). Since the clustering was performed with the optimal number of sea-

sons, which is two, this gives value to the hypothesis that the fragmentation in HKK Winter and Autumn is caused by a bad choice of the number of seasons.

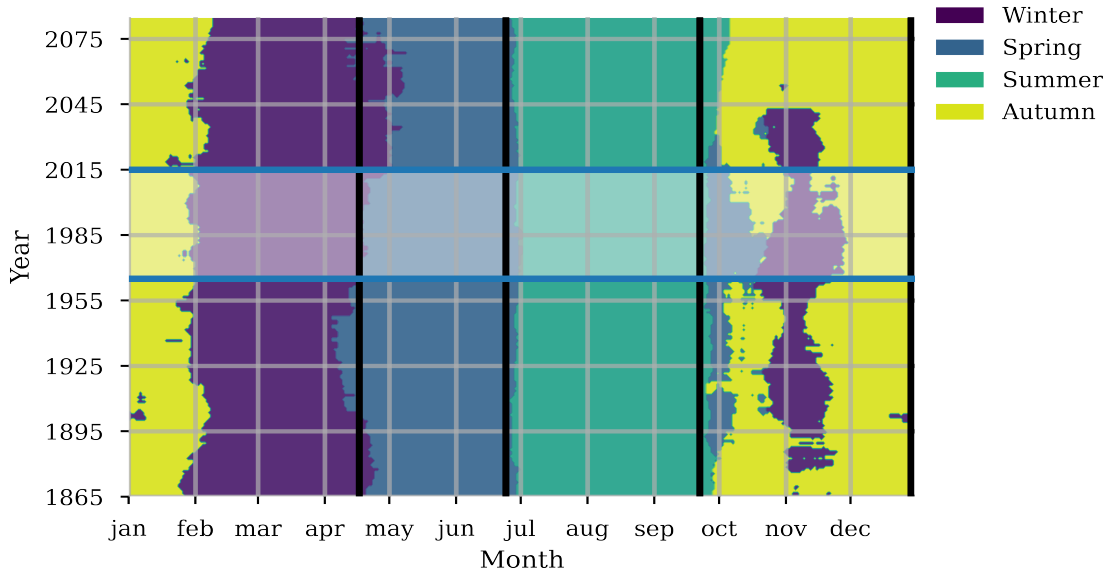

Figure 5.21: **Time evolution of seasons in HKK**: results of the SoftMax perceptron classification for the whole period in EC-Earth (1850-2015 historical, 2015-2100 SSP5-8.5 scenario). Black lines are the algorithm breakpoints, and white shaded area represents the period used in ERA5 for the seasons definitions.

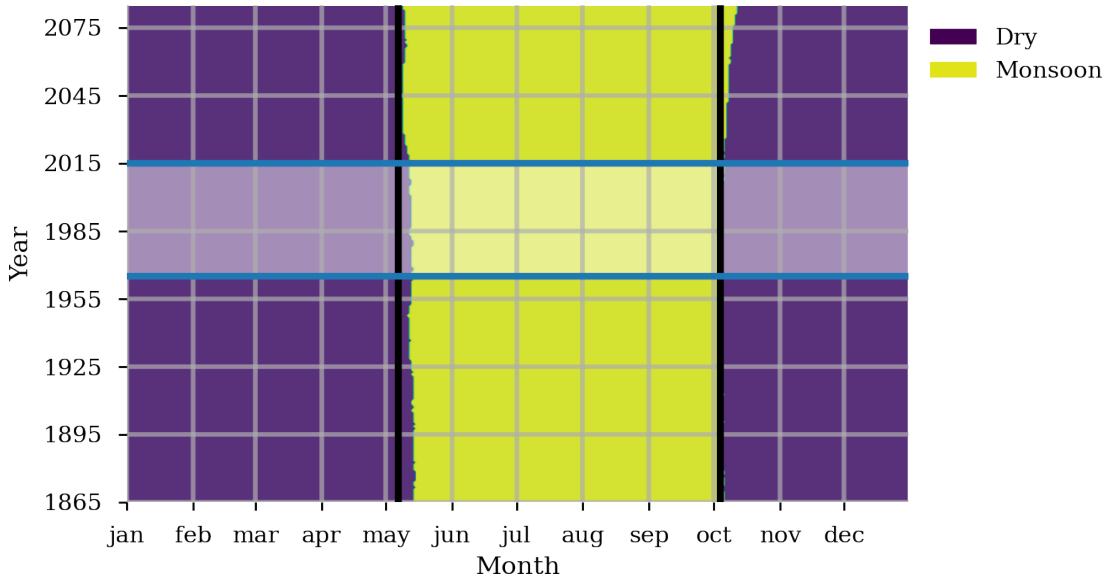

Figure 5.22: **Time evolution of seasons in Him**: same as Figure 5.21.

The next step is evaluating the time evolution of the length of each season. This is addressed simply counting the number of days contained in each season. We perform this operation on the results obtained for each ensemble member and for

the ensemble mean of probabilities obtained with the procedure explained in section 5.4.4.

In both regions we can note an increase in the Monsoon associated seasons length (named as Summer in HKK, Monsoon in Him) (Figures 5.23 and 5.24). While this is obviously related to a shortening of the dry season in Him (Figure 5.24), being only two seasons, this is not in HKK where we used four seasons. In HKK (Figure 5.23) we can note that Summer and Spring have a low interannual variability. They are stable in the historical period while in the future projection Summer tends to increase its duration, and Spring tends to shorten it. Winter and Autumn have a large interannual variability in the historical period, and also shows a large spread between the ensemble members. In this period they also show complementarity, i.e., when Autumn is longer Winter is shorter and vice versa. This confirms the difficulty of SoftMax perceptron in the identification of these two seasons in HKK, as stated above. Nevertheless it is worth to note that this behavior tends to disappear in the future projection under the SSP5-8.5 scenario. What we observe here is an increase in the duration of the Autumn season, and a shortening of Winter. There is a double interpretation of that: 1) the changes in seasonal patterns driven by the climate change are differentiating these two seasons making them more distinguishable for the SoftMax perceptron or 2) it is only a phase of a cycle, and if we could see further in time, we would see a behavior similar to the one observed in the historical period. Even in this case, an analysis performed with a different number of seasons might help answer this question.

The trend observed in Him is more clear (Figure 5.24), the shortening of the Dry season and the stretching of the Monsoon season is present in all the ensemble members, which also show a low spread.

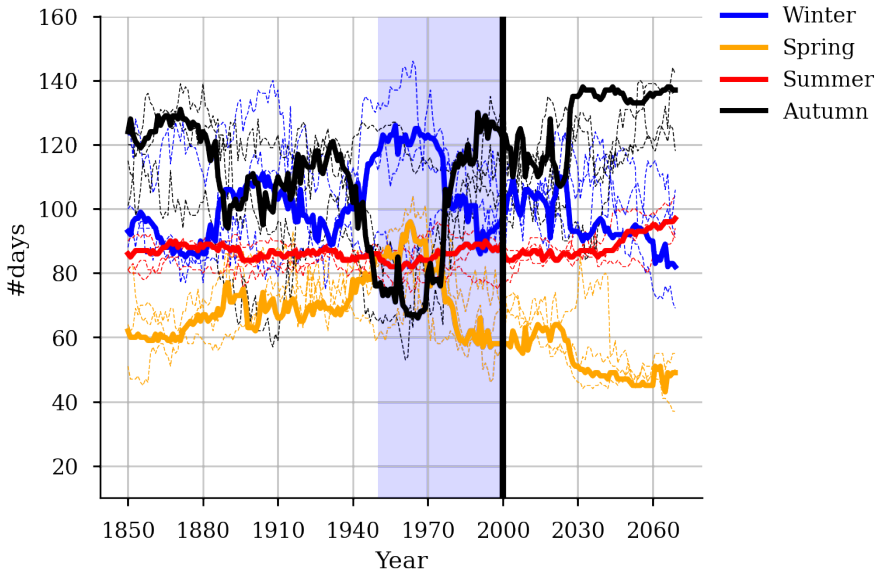

Figure 5.23: **Time evolution of seasons length in HKK**: dotted lines are the results of each ensemble member in Ec-Eart3h, while solid lines are the members' average.

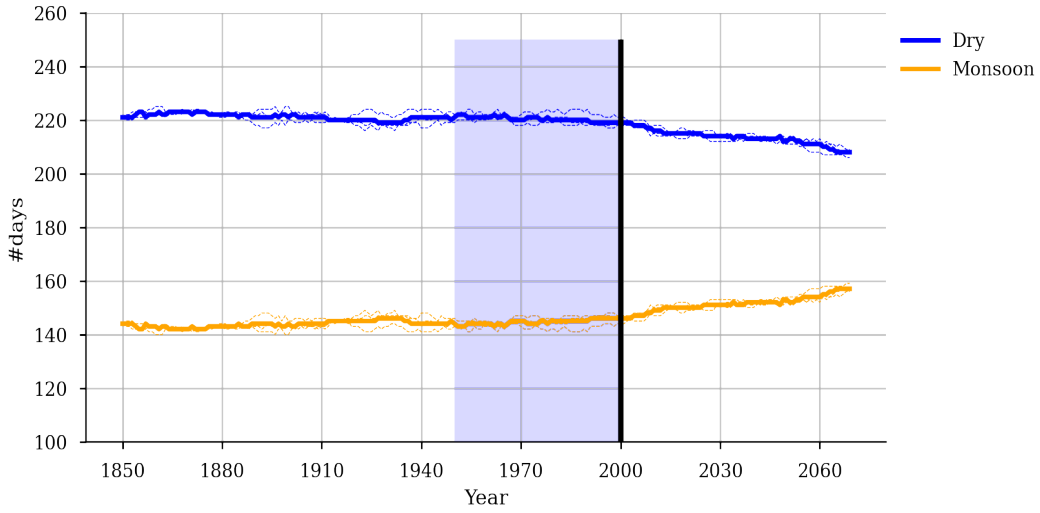

Figure 5.24: **Time evolution of seasons length in Him**: same as Figure (Figure 5.23).

5.4.5 Future trends with dynamical seasons

As we stated in section 5.3.3, a redefinition of seasonal breakpoints could imply a change in the mean and cumulative values of precipitation in the seasons, and therefore a change in their trends. Furthermore, using the dynamic breakpoints obtained from the SoftMax perceptron on EC-Earth3, the season length is no longer forced to be constant through years. This means that there could not be a linear relationship between cumulative and mean values trends, as happened with reference breakpoints. In this subsection we present the future trends on rainfall seasons mean and cumulative values are computed, computed using the dynamical breakpoints obtained before.

In HKK in all four seasons an increase in mean precipitation is expected, mainly in center and North areas (Figure 5.25 A, B, C, D), which are the mountain areas. This increase is lower than the one obtained with reference breakpoints in Summer (Figure 5.27 C) but with less spread between the grid points. In Winter (Figure 5.27 A) the mean values trends computed with dynamical breakpoints are slightly higher than the ones computed with reference breakpoints. Cumulative values trends computed with dynamical seasons are everywhere positive and higher than the ones obtained with reference breakpoints (Figure 5.27 E, F, G, H). This is particularly remarkable in Spring and Autumn (Figure 5.27 F H), which are the seasons which receive less precipitation. Spatially, this increase is concentrated in the North-West area in Winter (Figure 5.26 A), in the Center area in Spring (Figure 5.26 B), in the East and South areas in Summer (Figure 5.26 C) and in the South-West in Autumn (Figure 5.26 D)

A similar pattern could be observed in Him Box. Average values trends are higher in Winter (Figure 5.30 A) and lower in Summer (Figure 5.30 B) than the ones computed with reference breakpoints. On the other hand, cumulative values trends obtained with dynamical seasons are higher than the ones obtained with reference breakpoints (Figure 5.30 C, D). The increase in mean values is localized in the Center mountain area (Figure 5.25 A B). About Figure 5.28, we stated that with reference breakpoints there was a spot of significative high values in the South-East area, and we warned that could be caused by a stretching of the Monsoon season. This stretching was actually observed (Figures 5.22 and 5.24), and the trends are now positive not only in the South-East but also in the Center. In Him we can also note that the spatial distribution of cumulative values trends is similar to the one of mean values trends (Figure 5.28).

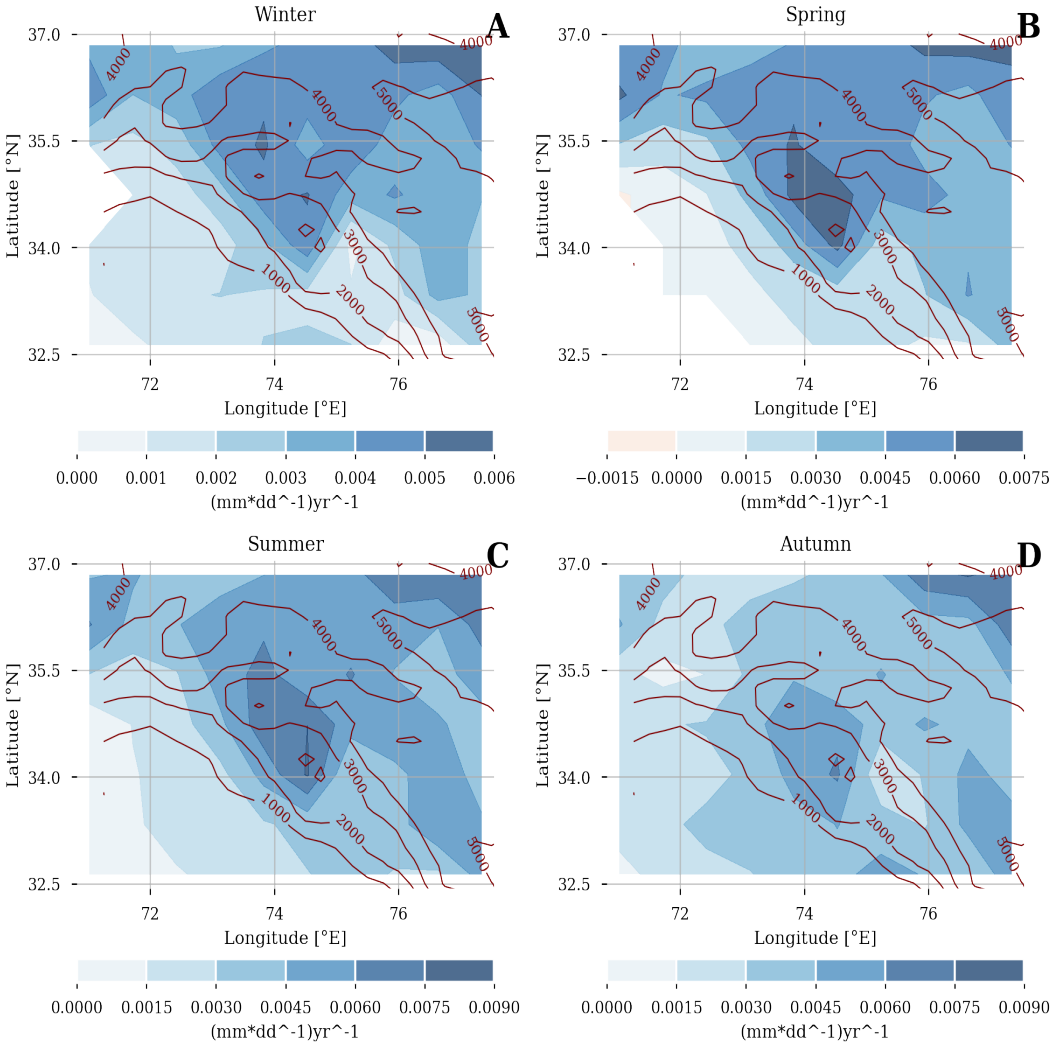

Figure 5.25: **HKK seasonal mean precipitation future trend maps (dynamical breakpoints)**: the method is the same as Figure 5.10 but with dynamical breakpoints.

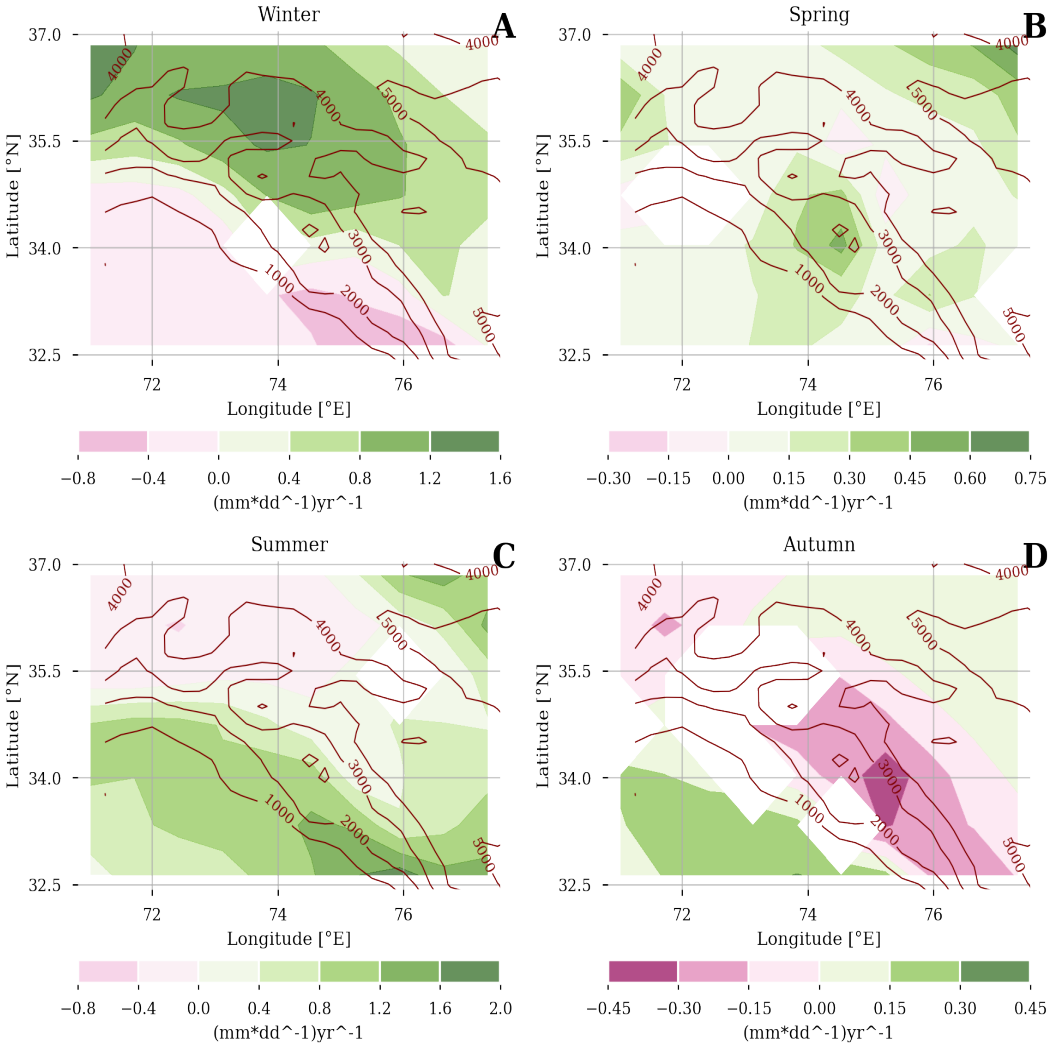

Figure 5.26: **HKK seasonal cumulative precipitation future trend maps (dynamical breakpoints)**: the method is the same as Figure 5.10 but with dynamical breakpoints and trends computed for cumulative values.

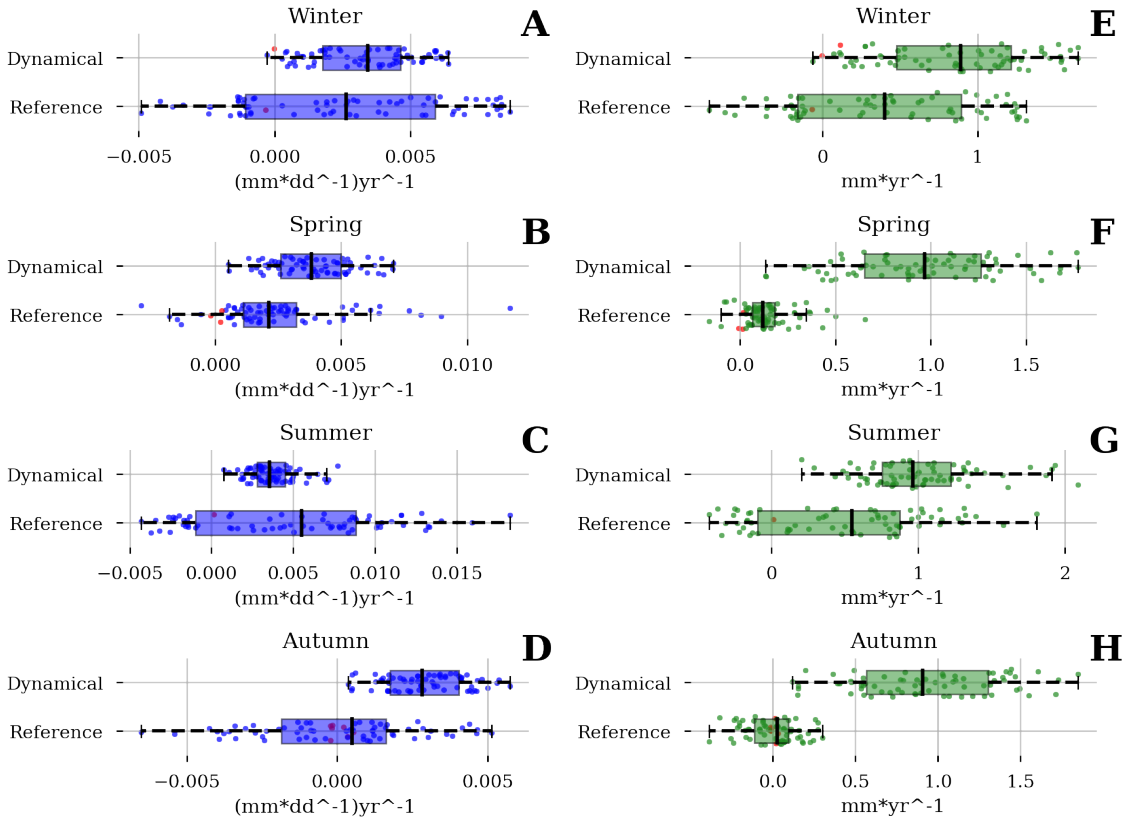

Figure 5.27: **HKK seasonal mean and cumulative future trend boxplots (dynamical and reference breakpoints)**: the method is the same as Figure 5.9 but with dynamical breakpoints and reference breakpoints for comparison (see y-axis labels). Letters A to D represents values and boxplots for mean values trends, letters E to H for cumulative values trends

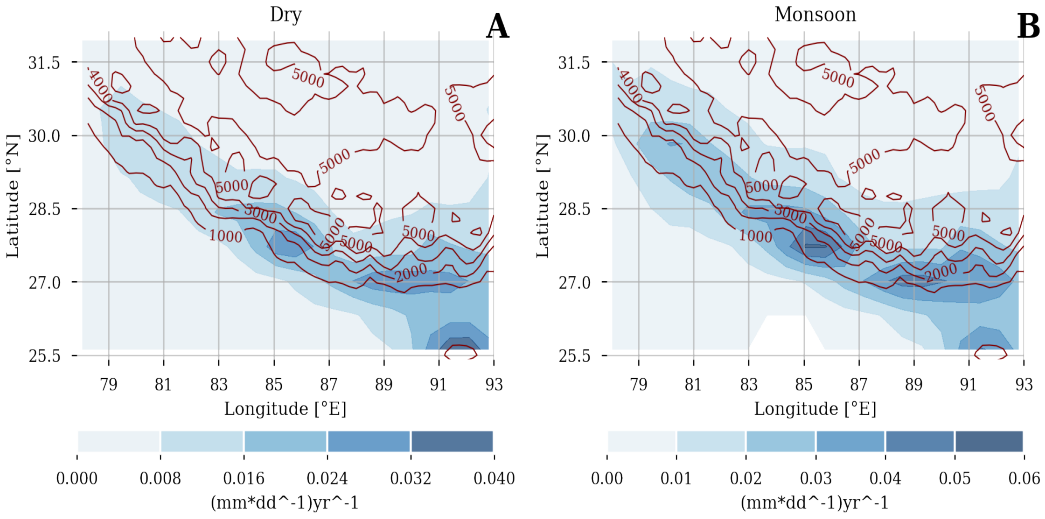

Figure 5.28: **Him seasonal mean precipitation future trend maps (dynamical breakpoints)**: the method is the same as Figure 5.10 but with dynamical breakpoints.

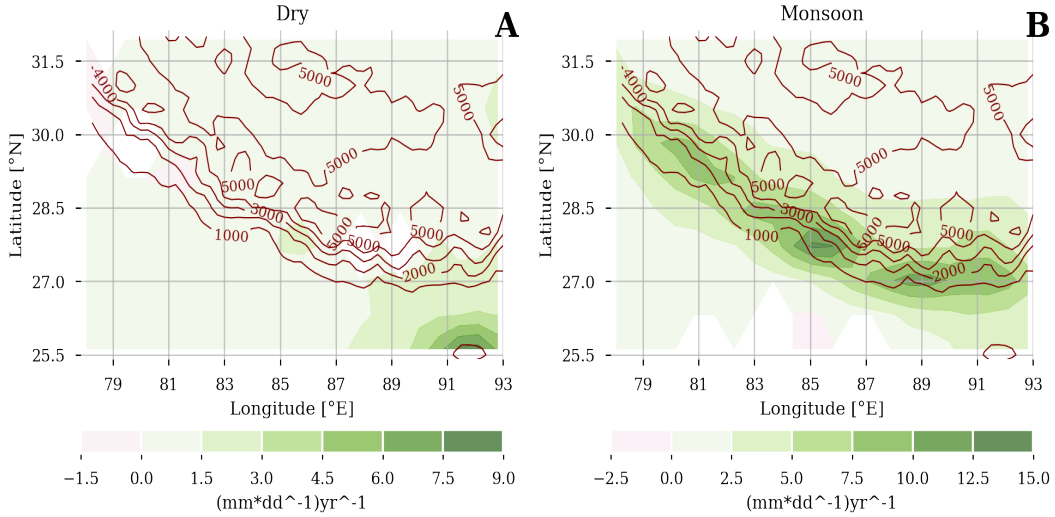

Figure 5.29: **Him seasonal cumulative precipitation future trend maps (dynamical breakpoints)**: the method is the same as Figure 5.10 but with dynamical breakpoints and trends computed for cumulative values.

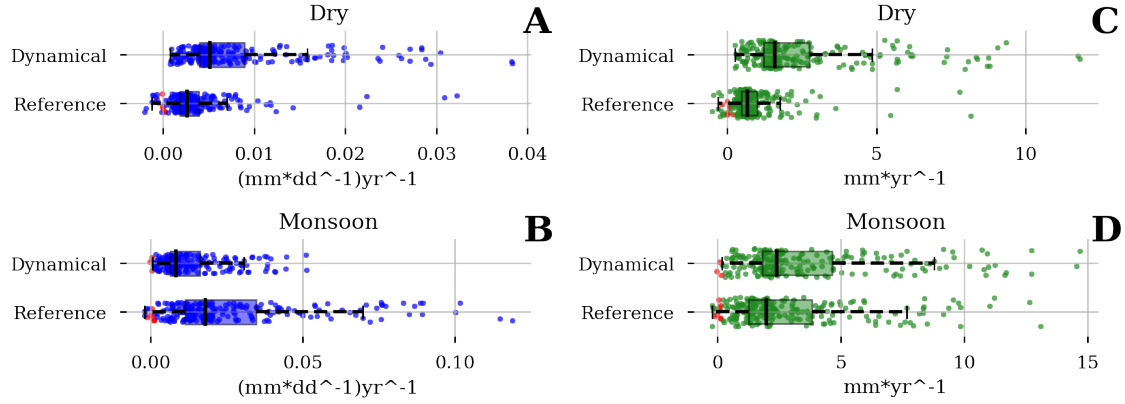

Figure 5.30: **HKK seasonal mean and cumulative future trend boxplots (dynamical and reference breakpoints)**: the method is the same as Figure 5.27. Letters A and B represents values and boxplots for mean values trends, letters C and D for cumulative values trends

5.5 Discussion

In this chapter we applied a set of data-drive approached designed for the definition (Radially Constrained Clustering) and classification (SoftMax perceptron) of the meteorological seasons to the region of Hindu Kush Karakoram Himalaya (HKKH), in present day and future climate, making use of both reanalysis and EC-Earth3 climate model data. Due to the peculiarity of the precipitation pattern in this region, we were interested in the evaluation of the seasonal cycle of precipitation.

A review on literature led us to the identification of the main dynamical mechanisms driving the precipitation in the area, which are the Indian Summer Monsoon (ISM) and the Western Disturbances (WDs). We referred to literature for the detection of the regions in which these phenomena are most relevant, so that we divided the HKKH region in the subregions of Hindu-Kush Karakoram (HKK) and Himalaya (Him). HKK is below the influence of both ISM and WDs, and shows a bimodal seasonal precipitation cycle with a peak in Winter and one in Summer. On the other hand, the seasonal cycle of precipitation over the Him sector is charctarized by only one peak in Summer, since it is not reached by WDs. In section 5.3.1 we showed the main features of these regions are respected in the dataset we used in this work, ERA5 and EC-Earth3, although some signficant bis exist in the Autumn precipitation pattern in the climate model.

Thus we analyzed the breakpoints, i.e., the days that mark the transition from a season to another. Firstly we referred to literature for the identification of the reference breakpoints. It results that HKK can be mainly described with four seasons (which we called Winter, Spring, Summer and Autumn in analogy with mid-latitude seasons) and Him with two seasons (usually defined as Dry and Monsoon). The breakpoints referring to the monsoonal season (Monsoon in Him and Summer in HKK) are the best assessed in literature, while the other seasonal transitions are not supported by the same amount of literature. We showed that, even though two is the optimal number of seasons for Him, in HKK a four seasons analysis from a data-driven point of view does not seem to be an optimal choiche (section 5.4.1): indeed, this leads to the identification of seasons which are not well differentiated. This is a remarkable point: in the continuation of the work we stated that a suboptimal choice in the number of seasons could badly affect the analysis. This ill-conditioning is not only a formal issue. In fact it may entail results that are difficult to interpret, leading to a bad understanding of the seasonal cycle. Thus, we point out that a more accurate choice of the number of the meteorological seasons is necessary in seasonal analysis, which are largely used in Climate sciences. Nevertheless, in order to get results comparable with literature we continued to use 4 seasons in HKK.

We thus evaluated the algorithm breakpoints in both regions with RCC on ERA5 data, and we found results relatively agreeing with reference breakpoints. This results underlines the adequacy of both the chosen algorithm (RCC) and metric (Euclidean distance). Indeed, they are capable of reproducing the results found in literature, which are calibrated to the specific regions, without needing to defining anything rather than the number of seasons. Nevertheless, a more accurate analysis of the implication of the choice of the metric should be carried out in followup work.

Then the seasons found with RCC in ERA5 have been used to train the SoftMax perceptron, which was used to classify the EC-Earth3 dataset. The training and testing of theSoftMax perceptron showed that it is capable of learning the seasonal features of the data, and detecting them in fresh new data. Thus, the investigation of different methods does not seem necessary.

Once the SoftMax perceptron has been trained, it has been used to classify the EC-Earth3 dataset. An overall analysis of the result shows that, while the classification in the Him box is stable, in HKK Autumn and Winter are fragmented. At the light of what we said before, this could be caused by two reasons: 1) the bias in the EC-Earth3 representation of the seasonal precipitation pattern, with a third peak in November, 2) the suboptimal choice of the number of seasons. The first hypothesis is supported by the fact that a spot of Winter days is found in November, and the second by the fact that the fragmentation is also present at the seasonal boundaries, suggesting that seasons are not well separated as they appear to be for example in Spring. This fragmentation tends to shrink in future projection under SSP5-8.5 scenario. Thus we assumed that the change in seasonal cycle driven by Climate Change is increasing the differences between these two seasons. This application shows that the methodology we developed can be used as a tool for the validation of the seasonal cycle representation in Earth systems datasets.

In the last part of this chapter, we applied the classification performed with SoftMax for the identification of the dynamical breakpoints. Thus, we used the dynamical breakpoints for the evaluation of seasons length and for the computation of the future trends under SSP5-8.5 scenario of precipitation seasonal mean and cumulated values. We thus compared these results with the ones obtained with the reference breakpoints. About these results, an overall increase in precipitation amount is reported with both seasons definitions. This is not surprising, since global increase in precipitation in all seasons is something well assessed in literature and in line with what theoretically expected from a global warming perspective. The most remarkable result we obtained with the introduction of dynamical seasons is the correction in the intensity of these trends. The seasons associated with the Indian Summer Monsoon (Summer in HKK and Monsoon in Him) are probably going to

increase in duration. This implies that mean precipitation values computed with dynamical seasons are lower than the ones obtained with reference breakpoints. On the other hand, cumulative values are higher. The opposite is expected to occur in the Winter season in HKK and in the Dry season in Him. These seasons are going to decrease in duration, with mean values computed with dynamical breakpoints higher than the ones computed with reference breakpoints, and cumulative values lower than the ones computed with reference breakpoints. Once again, this has not only formal implications. The power of seasonal analysis consists in the fact that they can condense information about the state and evolution of climate, which are used, among other things, for the development of strategies of adaptation. A more correct and complete extraction of this information could lead to a better understanding of the climate system.

Chapter 6

Conclusions

We started this thesis wondering if the meteorological seasons could be defined in a way which is more robust and reliable than the heuristic approach which is commonly used nowadays. This led to the work hypothesis on which the whole work has been based on, that is to say that seasonality in Earth's climate system leads to the emergence of periods within a year with similar statistical behaviour, which are internally similar, and well differentiated from each other. We also stated that this assumption is non trivial, since a continuous periodical signal such as seasonality could not be suitable for this division. In light of the results obtained in the case study, we can conclude that our work hypotesis was well conditioned. The analysis performed on the Hindu-Kush Karakoram/Himalaya (HKKH) regions resulted in the recognition of seasons which are physically meaningful, being able to reproduce the main physical features of the seasonal patterns that we identified in a review on literature.

For the practical recognition of the meteorological seasons in climate data, we decided to rely on a set of machine learning tools. The choice of machine learning was driven by the fact that these kind of algorithms are being deployed in a wide range of applications, included climate sciences, with remarkable results. In our case, we expected to obtain from machine learning algorithms a recognition of the meteorological seasons which is physically meaningful and well understandable, without the need of human supervision. We selected two algorithms: the Radially Constrained Clustering (RCC) for the recognition of seasons, and the SoftMax perceptron for the evaluation of their evolution in different periods and/or datasets.

With the application of these algorithms to the HKKH case study, we obtained multiple interesting results. Firstly, the evaluation metrics for the RCC on ERA5 suggested a correction for the total number of seasons to use for the description of the seasonal patterns of total precipitation and surface air temperature, with respect to the number used in our literature references. Forcing the clustering to

perform using the number of seasons found in literature resulted in a definition of meteorological seasons which is similar to the reference one. We can thus conclude that the RCC is a good tool for the definition of the meteorological seasons, and that the metric we chose (Euclidean distance) is suitable for this purpose.

Then we used the SoftMax perceptron for evaluating the evolution of these seasons in the global climate model EC-Earth3. We focused on how this dataset reproduce the seasons with respect to ERA5, and how these seasons will change in the future, under the SSP5-8.5 scenario. Here we found that the SoftMax perceptron is able to identify a bias in the seasonal pattern of precipitation which is well assessed in literature. For the evolution in the future, we evaluated the length of the seasons and we compared the trends of two seasonal metrics (daily average precipitation and seasonal cumulated precipitation), computed with both the seasons found in literature and the ones obtained with the SoftMax perceptron. We found that the length of the seasons in HKKH is expected to change considerably in the future, and that this comports some corrections on the trends we evaluated. We do not have enough elements to state if these correction are correct, and we leave these kind of evaluation for followup works. On the other hand, we can conclude that approaching the division in seasons using 'static' seasons, that is to say a division which is the same through different periods, is a wrong choice, especially with the changes that are projected in future climate due to climate change.

We can conclude this thesis stating that a more robust and rigorous approach to the meteorological seasons with respect to the one which is commonly used nowadays is possible. Meteorological seasons are entities which can be defined in several ways depending on the variables and locations we take into account, and their recognition through specific, physically-driven approaches, could result in an onerous work. The data-driven way using machine learning algorithms has proven to be reliable, easy to implement, and able to give a better understanding of what meteorological seasons are and how they will change in the future.

List of Figures

List of Tables

Bibliography

[Amrith, 2018] Amrith, S. (2018). Risk and the south asian monsoon. *Climatic Change*, 151:17–28.

[Bingyi, 2005] Bingyi, W. (2005). Weakening in indian summer monsoon in recent decades. *Advances in atmospheric sciences*, 22:21–29.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

[Cai et al., 2021] Cai, S., Wang, Z., Wang, S., Perdikaris, P., and Karniadakis, G. (2021). Physics-informed neuralnetworks for heat transfer problems. *Journal of Heat Transfer*, 143.

[Cannon, 2005] Cannon, A. J. (2005). Defining climatological seasons using radially constrained clustering. *GEOPHYSICAL RESEARCH LETTERS*, 32.

[Chen et al., 2018] Chen, R., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations.

[Deepa and Oh, 2014] Deepa, R. and Oh, H. O. (2014). Indian summer monsoon onset vortex formation during recent decades. *Theoretical and Applied Climatolology*, 118:237–249.

[Dimri et al., 2016] Dimri, A. P., Niyogi, D., Barros, A. P., Ridley, J., Mohanty, U., Yasunari, T., and Sikka, D. R. (2016). Western disturbances a review. *Rev. Geophys*, 53:225–246.

[Documentation, 2018] Documentation, E. S. (2018). Cmip6 tier 1 experiment: historical.

[Döscher et al., 2022] Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzik, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X.,

and Zhang, Q. (2022). The ec-earth3 earth system model for the coupled model intercomparison project 6. *Geoscientific Model Development*, 15(7):2973–3020.

[ECMWF, 2023] ECMWF (2023). Copernicus Climate Data Store.

[ESGF, 2023] ESGF (2023). Earth system grid federation.

[Eyring et al., 2016] Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

[Gadgil, 2003] Gadgil, S. (2003). The indian monsoon and its variability. *Annual Review of Earth and Planetary Sciences*, 31:429–467.

[Hamilton, 1994] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton New Jersey United Kingdom.

[Hersbach et al., 2020] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

[IPCC, 2021a] IPCC (2021a). Ipcc special report on the ocean and cryosphere in a changing climate, annex v: The monsoon system. *Climate Change 2021 The Physical Science Basis Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 2193–2204.

[IPCC, 2021b] IPCC (2021b). Linking global to regional climate change. *Climate Change 2021 The Physical Science Basis Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1363–1512.

[Katzenberger et al., 2021] Katzenberger, A., Schewe, J., Pongratz, J., and Levermann, A. (2021). Robust increase of indian monsoon rainfall and its variability under future warming in cmip6 models. *Earth System Dynamics*, 12:367–386.

[Kripalani et al., 2003] Kripalani, H. R., Kulkarni, A., and Sabade, S. S. (2003). Indian monsoon variability in a global warming scenario. *Natural Hazards*, 29:189–206.

[Krishnan, 2019] Krishnan, R. (2019). Unravelling climate change in the hindu kush himalaya and increasing extremes. *The Hindu Kush Himalaya Assessment: Mountains, Climate Change, Sustainability and People*, pages 57–97.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Liu et al., 2015] Liu, B., Liu, Y., Wu, G., Yan, J., He, J., and S, R. (2015). Asian summer monsoon onset barrier and its formation mechanism. *Clim Dyn*, 45:711–726.

[Lloyd, 1957] Lloyd, S. (1957). A least squares algorithm for fitting surfaces. *IEEE Transactions on Information Theory*, IT-3(4):227–231.

[Lugosi and Cesa-Bianchi, 2005] Lugosi, G. and Cesa-Bianchi, N. (2005). The cross entropy method for classification. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML)*, page 472–479.

[Mahesh, 2018] Mahesh, B. (2018). Machine learning algorithms - a review.

[Mao and Wu, 2007] Mao, J. and Wu, G. (2007). Interannual variability in the onset of the summer monsoon over the eastern bay of bengal. *Theoretical and Applied Climatolology*, 89:150–170.

[Minsky and Papert, 1969] Minsky, M. and Papert, S. (1969). *Perceptrons: an introduction to computational geometry*. MIT press.

[NIST, 2012] NIST (2012). NIST/SEMATECH e-Handbook of Statistical Methods.

[Novikoff, 1962] Novikoff, A. (1962). On convergence proofs for perceptrons. *Proceedings of the Symposium on the Mathematical Theory of Automata*, pages 615–622.

[O'Neill et al., 2016] O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., and Sanderson, B. M. (2016). The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482.

[Overpeck et al., 2011] Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R. (2011). Climate data challenges in the 21st century. *Science*, 331.

[Pai et al., 2020] Pai, D. S., Bandgar, A., Davi, S., Musale, M., Badwaik, M. R., Kundale, A. P., Gadgil, S., Mohapatra, M., and Rajeevan, M. N. (2020). Normal dates of onset/progress and withdrawal of southwest monsoon over india. *MAUSAM*, 71:553–570.

[Pai and Rajeevan, 2009] Pai, D. S. and Rajeevan, M. N. (2009). Summer monsoon onset over kerala: New definition and prediction. *Journal of Earth System Sciences*, 118:123–135.

[Palazzi et al., 2013] Palazzi, E., Von Hardenberg, J., and Provenzale, A. (2013). Precipitation in the hindu-kush karakoram himalaya observations and future scenarios. *Journal of geophysical research*, 118:85–100.

[Palazzi et al., 2015] Palazzi, E., Von Hardenberg, J., Provenzale, A., and Terzago, S. (2015). Precipitation in the karakoram-himalaya: a cmip5 view. *Clim Dyn*, 45:21–45.

[Pletzer and Hayek, 2018] Pletzer, A. and Hayek, W. (2018). Mimetic interpolation of vector fields on arakawa c/d grids. *Journal of Applied Meteorology and Climatology*.

[Reichstein et al., 2019] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204.

[Rosenblatt, 1959] Rosenblatt, F. (1959). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

[Roxy et al., 2015] Roxy, M. K., Ritika, K., Terray, P., Murtugudde, R., Ashok, K., and Goswami, B. N. (2015). Drying of indian subcontinent by rapid indian ocean warming and a weakening land-sea thermal gradient. *Nature communications*.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Seyed et al., 2006] Seyed, F. S., Giorgi, F., Pal, J. S., and King, M. P. (2006). Effect of remote forcings on the winter precipitation on central southwest asia part 1: observations. *Theoretical and Applied Climatolology*, 86:147–160.

[Steinbach et al., 2006] Steinbach, M., tab, P., Boriah, S., and Kumar, V. (2006). The application of clustering to earth science data:progress and challenges.

[Taylor et al., 2018] Taylor, E. K., Martin, J., V., B., Luca, C., Sébastien, D., Paul, J. D., Mark, E., Eric, G., Slava, K., Michae, L., Bryan, L., Denis, N., and Stockhause, M. (2018). Cmip6 global attributes drs filenames directory structure and cvs.

[Tsypkin, 1968] Tsypkin, Y. Z. (1968). Adaptive stabilization of systems with several unknown parameters. *Automation and Remote Control*, 29(5):670–679.

[Ward Jr and Hooker, 1963] Ward Jr, J. H. and Hooker, J. N. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[Widrow and Stearns, 1990] Widrow, B. and Stearns, S. D. (1990). The adaptive noise cancelling algorithm. *Proceedings of the IEEE*, 63(12):1692–1716.

[Wu and Zhang, 1998] Wu, G. and Zhang, Y. (1998). Tibetan plateau forcing and the timing of the monsoon onset over south asia and the south china sea.

[Xu and Wunsch, 2015] Xu, R. and Wunsch, D. (2015). A survey of clustering algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 26(1):11–30.

[Xu and Rutledge, 2019] Xu, W. and Rutledge, S. A. (2019). Time scales of shallow-to-deep convective transition associated with the onsetof madden-julian oscillations. *Geophysical Research Letters*, 43.

[Yuan and Yang, 2019] Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235.

[Zhisheng et al., 2015] Zhisheng, A., Guoxiong, W., Jianping, L., Youbin, S., Yimin, L., Weijian, Z., Yanjun, C., Anmin, D., Li, L., Jiangyu, M., Hai, C., Zhengguo, S., Liangcheng, T., Hong, Y., Hong, A., Hong, C., , and Juan, F. (2015). Global monsoon dynamics and climate change. *Annual Review of Earth and Planetary Sciences*, 43:2.1–2.49.