Universität Leipzig

Fakultät für Mathematik und Informatik
Institut für Informatik

Analyzing the Evolution of Moisture Transport Patterns in the North Atlantic based on Ensemble Simulations

Masterarbeit

Leipzig, August 2024

vorgelegt von

Denis Streitmatter
Studiengang Master Informatik

Betreuende Hochschullehrer:
Prof. Dr. Gerik Scheuermann
Universität Leipzig, Abteilung für Bild und Signalverarbeitung

Abstract

The distribution and variability of precipitation in Europe are significantly influenced by moisture transport over the north(east)ern Atlantic. Due to the turbulent nature of moisture transport, structural changes are difficult to track, which is tackled by analyzing the main variability patterns with a sliding window approach. The objective of this thesis is to visually analyze and compare changes in different future climate scenarios. In addition, we investigated connections with dominant Atlantic oscillation patterns (North Atlantic Oscillation (NAO) and East Atlantic Pattern (EAP)) and precipitation in Europe. Based on the latest Max Planck Institute Grand Ensemble CMIP6, visualizing the variability introduced by the 50 members of the simulation poses a challenge. To mitigate the visual clutter associated with the representation of multiple members' contour lines, a hexbin-based approach was used to facilitate the analysis of variability introduced by these numerous members. The results identified two dominant modes of water vapor transport, which demonstrated considerable stability across different members and time periods, along with structural changes in several spatial patterns. In general, the variability explained by all moisture transport variability patterns increases, especially in pronounced climate change scenarios. This effect was also observed in the primary pattern of precipitation and EAP. The modes of moisture transport also exhibited significant correlations with the leading oscillation and precipitation patterns.

Contents

1 Introduction and Motivation

1.1 Motivation

Since the discovery (and further confirmation) of the greenhouse effect in the years from 1824 to 1900 [Foo56; Fou24], mankind has come a long way in understanding the consequences of the increased concentration of greenhouse gases in the atmosphere of the Earth. Especially in the last decades, the climate crisis has gained more and more attention, leading to the creation of several international organizations and institutions (e.g., the Intergovernmental Panel on Climate Change (IPCC) in 1988). In 2019, more than 11,000 scientists around the world issued a statement [Rip+19] calling on governments around the world to take action.

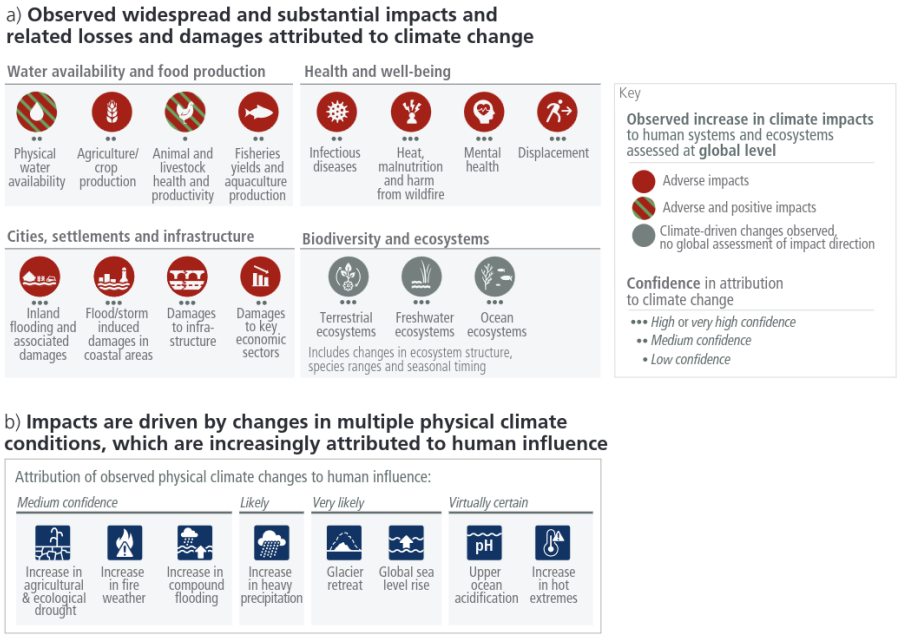


Figure 1.1: Impact of Climate Change for Humans, taken from the 6th IPCC report for policymakers [Lee+24]

The consequences for the environment and humans are prevalent and are, in part, already visible today. Figure 1.1 shows the likely consequences for humans from the latest IPCC report for policymakers [Lee+24]: Flooding, malnutrition, displacement, and ecosystem damage can be attributed with great confidence to climate change.

Atmospheric moisture and precipitation are main part of the water cycle and provide large parts of the water consumed by plants, crops, animals, and humans. The variability of precipitation, and its main source, atmospheric moisture transport, are drivers of floods and droughts, and they play a vital role for humans and ecological systems. The variability in these processes is significantly influenced by dominant oscillation systems, such as NAO in Europe and the El Niño Southern Oscillation (ENSO) in South America. Recent research shows that big circulation systems like the NAO [Vie+21a] change as well in the face of climate change, depending on its intensity. This thesis aims to investigate the implications of these changes specifically in the context of moisture transport. Moisture transport, a critical component of the atmospheric system, presents significant challenges for analysis due to its inherently oscillating nature. To address the complexity of moisture transport, the thesis attempts to tackle this issue by employing pattern analysis to track structural changes of moisture and a sliding window approach (motivated by research of [Vie+21a]) to identify and understand structural changes in moisture transport mechanisms in Europe and the Northern Atlantic. Furthermore, reducing the turbulent flow of moisture to its dominant patterns helps address one of the major challenges in modern climate science: the uncertainty associated with multiple simulation results and the reduction of the immense volume of data.

1.2 Climate and Climate Research

1.2.1 Quick Overview over Climate Systems and Climate Change

While weather is the momentary state of the atmosphere at a time, climate is the average of weather patterns over a longer period of time, usually 30 years or more [NOA]. So, the term climate change does not refer to any unexpected weather changes, but to the structural changes of said patterns over a large period of time (e.g., the warming of the global average temperature). The climate system of the Earth can be seen as complex interactions of its main components: atmosphere, hydrosphere, cryosphere, lithosphere, and biosphere [Int23; Vie24]. Changes in this system can have (roughly) two reasons: Either redistributions of energy, presenting themselves as internal oscillations, which can happen

on arbitrary large scales[1] or in the form of external forcings [Vie24]. External forcings influence the system from external sources, such as volcanic activity, variations in solar radiation, and, most importantly, the emission of Greenhouse Gasses (GHG).

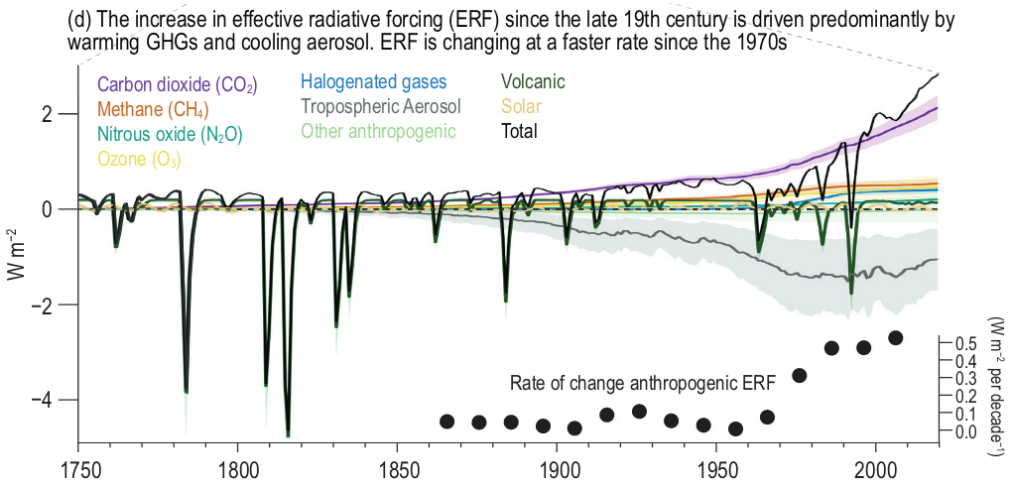


Figure 1.2: The evolution of the effective radiative forcing and contributing components, taken from [Int23]. The black line indicates the total Effective Radiative Forcing (ERF), while the contributions of different forcings are colored. Black dots at the bottom indicate the rate of anthropogenic Effective Radiative Forcing (ERF) change.

Figure 1.2 shows an example of an effect of such external forcings: It shows the change in Effective Radiative Forcing (ERF) and its contributing components over the last 2.5 centuries. ERF (in Wm^{-2}) is a way to measure how much energy from the sun is "trapped" instead of reflected back to space (greenhouse effect) and describes the energy balance in the atmosphere of the Earth. A positive value means warming, while a negative value is associated with cooling. As illustrated in Figure 1.2, there was no significant change in natural forcings (volcanic or solar radiation), the main drivers of change in ERF are clearly the man-made GHGs and cooling aerosols. [Int23]

Regarding the internal variations: Most of it is part of a cyclical, spatio-temporal pattern. Especially the interaction of the atmosphere with the hydrosphere (i.e. all liquid forms of water on Earth) are responsible for large parts of the climate's internal variations on decadal and interannual time frames [Vie24]. Prominent examples of such oscillations are ENSO or NAO, the latter being especially relevant for this thesis.

[1] See the discussion on the change of the Atlantic meridional overturning circulation in the work of Lobelle et al. [Lob+20], which could be a multidecade-spanning oscillation

1.2.2 The North Atlantic Oscillation

The aforementioned NAO is "one of the most recurrent and prominent patterns of atmospheric circulation variability" [Hur+03]. It is also one of the oldest known weather patterns, since descriptions of Scandinavians exist from centuries back. It dictates the climate variability for a large area: From the East Coast of the USA to Siberia and from the Arctic to the subtropical Atlantic. Especially in boreal winter (usually from December to February), the variations of the NAO influences a wide range of variability areas: From the mean wind speed and direction to the heat and moisture transport, as well as the intensity and amount of storms and their path. [Hur+03]

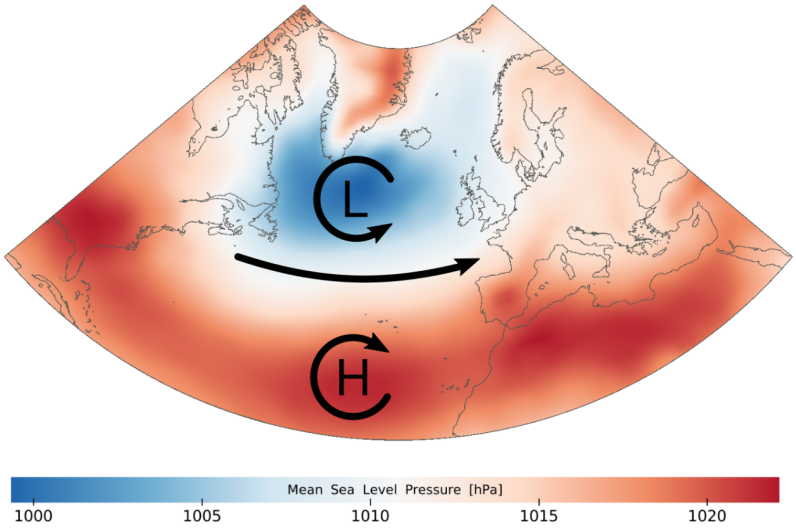


Figure 1.3: Characteristic mean SLP field of a boreal winter season, taken from [Vie24]. It shows a high (H)/low (L) pressure pattern, directing air from the Atlantic westwards towards Europe.

The NAO is a redistribution of atmospheric mass from the Arctic to the subtropical Atlantic, producing the aforementioned effects while swinging from one phase to another. Its basis is a characteristic dipole in the Sea Level Pressure (PSL) field of the Atlantic (see Figure 1.3). Due to the Coriolis force, air flows clockwise around high pressure and counterclockwise around low pressure in the Northern Hemisphere, leading to the transport of maritime air from the Atlantic towards Europe (see Figure 1.3) [Hur+03; Vie24]. Depending on the pressure differences, the effect varies: high pressure differences lead to higher transport of mild, humid air to Europe, which in turn results in milder European winters. In contrast, a low difference leads to a less pronounced effect and therefore to colder winters [Vie24]. These pressure differences vary on an interannual scale, and this effect is called the

North Atlantic Oscillation [Hur+03]. Figure 1.4 illustrates the NAO index, which is based on measurements of weather stations in Iceland and the Azores (top row), and represents the conventional methodology for defining the NAO. Another method is by computing the first/dominant Empirical Orthogonal Function (the pattern analysis technique employed in this thesis, see Section 2.2) of the sea level pressure field in wintry North Atlantic/Europe, the temporal coefficients (or Principal Components) (middle row) are a good estimate of the measured index.

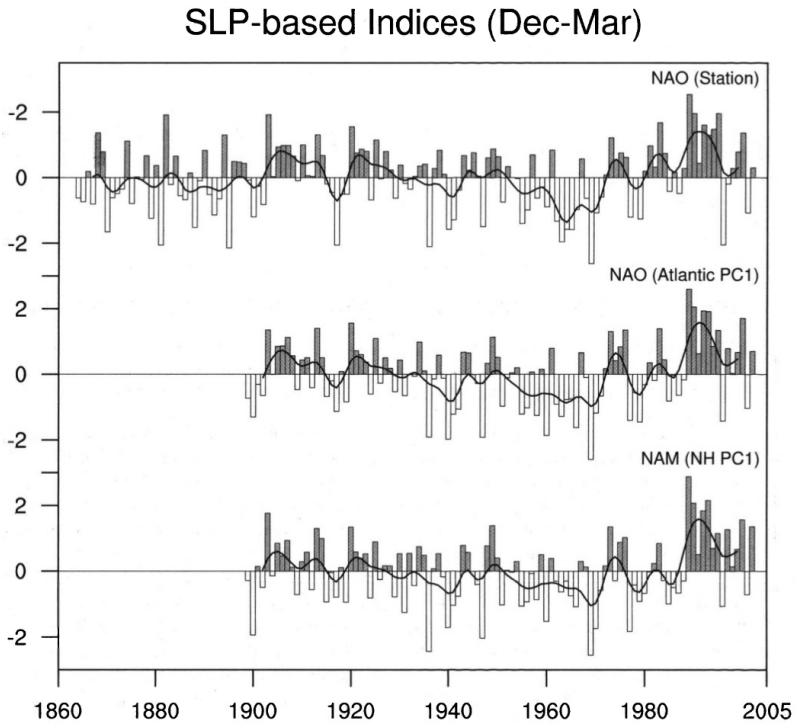


Figure 1.4: Comparison of the NAO index from [Hur+03]: The top panel are differences of SLP from weather stations in Portugal and Iceland, the middle panel is the first principal component of corresponding to the first Empirical Orthogonal Functions (EOF) of the northern Atlantic SLP field, and the bottom panel is the same as the middle panel but for the whole Northern Hemisphere. See [Hur+03] for a more detailed description

A large fraction of the recent warming in Europe can be linked to the behavior of the NAO in the last decades: it shifted from large amplitude anomalies in the negative direction to similar anomalies in the opposite direction in the later years. Therefore, [Hur+03] points out the need to study the relationship of anthropogenic climate change and the NAO. Following the argumentation of [Hur+03], the motivation of this thesis of Vietinghoff et al. [Vie+21a] was to track the shift of the centers of the dipoles in different climate scenarios.

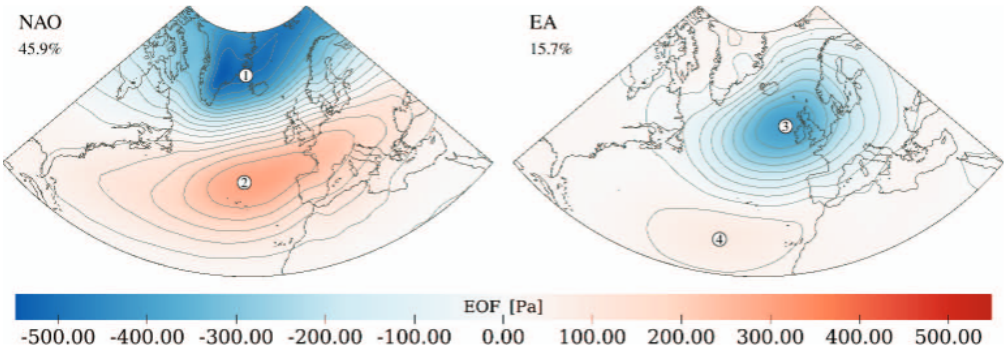


Figure 1.5: The most prominent modes of internal atmospheric variability in Europe/the Atlantic: The North Atlantic Oscillation (NAO) and East Atlantic Pattern (EAP) (taken from [Vie+21b])

Also worth mentioning is the East Atlantic Pattern (EAP, see Figure 1.5), the second most prominent mode of internal atmospheric variability. It is defined as the second Principal Component (Empirical Orthogonal Functions (EOF)) of PSL in the Atlantic. It is characterized by a large Sea Level Pressure Anomaly in the west of the British Islands and less pronounced anomalies in the subtropical Atlantic and eastern Europe. Similarly to the NAO, it is prominent throughout winter and influences precipitation and temperature throughout Europe. [CMW16; SLP19]

1.2.3 Climate Research: The IPCC and the Coupled Model Intercomparison Project (CMIP)

The rationale behind the UN General Assembly's endorsement of the IPCC in 1988 was to facilitate the preparation of comprehensive reviews and reports concerning the current state of scientific knowledge and research. Subsequently, six assessment cycles have been conducted, resulting in the publication of six reports that synthesize the findings of the scientific community. Figure 1.1, taken from the most recent report for policy makers [Lee+24], illustrates the potential impacts of climate change on humans.

The primary source for these figures in the reports are the so-called Global Coupled Models (GCMs), which attempt to model the state and evolution of specific fields of Earth data. These models are composed of multiple components, each representing a significant aspect of the Earth's intricate climate system, including the atmosphere and hydrosphere. In addition, they facilitate the examination of the dynamic interactions between these components [Vie24]. In the mid-1990s, the CMIP was established with the objective of enhancing the consistency and comparability of GCM results. CMIP provides the outer

structure for the production of simulations, the type of simulation (e.g. pre-industrial control simulations, future scenarios, etc.), the generation of fields, the provision of resolutions, and the serialization of results. Subsequently, the results of CMIP have assumed an increasingly prominent role in IPCC reports [TBL20], to the extent that they are now considered a fundamental element of climate science [Eyr+16]. The CMIP is currently in its sixth phase, corresponding to the recently published sixth assessment report of the IPCC [Lee+24]. The sixth phase describes an inner core (DECK[2] + historical simulations), which is a prerequisite for participation in CMIP. It also includes some endorsed Model Intercomparison Projects (MIPs), which are optional. Examples of these include ScenarioMIP (future scenario simulations), HighResMIP (for exploring models with higher resolutions), and GeoMIP (exploring the effects of geoengineering). [Eyr+16]

The simulations are usually set up as so-called ensemble simulations. This means that they consist of different members, which are one realization or run of a simulation. The members use the same forcings, but different starting conditions and are independent of each other. Using multiple simulations, it is possible to separate internal variability from responses to external forcing, enabling researchers to better quantify the consequences of climate change. Furthermore, it makes the research of extreme weather phenomena (e.g., droughts, floods, etc.) more robust in spite of their rare occurrences [MML21]. However, this approach also presents a challenge in terms of visualization, as it requires displaying multiple variants of the same data simultaneously, which can be difficult to do effectively. This issue was also identified as a significant research problem in the field of scientific visualization [Joh04].

1.3 Research Questions and Thesis Structure

Following up on the previous sections, the research question for this thesis is the following.

> "How do the Patterns of Moisture Transport and precipitation change in the face of different climate scenarios in the North-East Atlantic?"

Patterns are needed to reduce the sheer amount of data and make it possible to compare different climate scenarios across multiple members of the simulation ensemble beyond simple statistics. With this goal, the broad research question can be broken down into smaller milestones.

[2] Diagnostic, Evaluation and Characterization of Klima. This is a set of baseline simulations that should be included in every next CMIP iteration. An example of such an experiment is a pre-industrial control simulation.

M1: Generate patterns of moisture transport and other variables

Moisture transport must be quantified and this quantification must be calculated based on the data available in the chosen dataset (Chapter 3). Furthermore, a similar sliding window approach as in [Vie+21a] needs to be implemented to study the evolution of the patterns. The challenge is hereby the large size of the datasets due to multiple members, scenarios, and the combination of a large time scope and temporal resolution.

M2: Study the relationships with other variables and patterns

To grasp the meaning of moisture transport, the connection or relation to other variables (see Section 4.5) and patterns needs to be explored. The first connection should be to the NAO, or generally, patterns of surface pressure levels (such as the EAP the second most significant mode of PSL EOF). The second connection should be precipitation (patterns), as one of the most important consequences of transported moisture and a great influence on ecological and economic systems.

M3: Visualize the results

The patterns, its components, and their relationships with each other and variables need to be visualized so that they can be properly interpreted. It is important here to visualize the variability introduced by multiple members of the ensemble simulation. This includes selecting a feature in the results that can be used to analyze this variability and its change over time. For example, one interesting thing to find is if moisture transport experiences a similar shift to the north similar to the results of Vietinghoff et al. [Vie+21a].

The goal of this thesis is not to interpret the results, but rather to provide ideas, algorithms, and visualizations for climate scientists to interpret the results of changing moisture transport EOF patterns.

The remaining thesis is structured as follows: Chapter 2 introduces the theoretical background on fields and pattern analysis. The following Chapter 3 provides a detailed overview about the used CMIP6 based dataset. Chapter 4 provides an overview of the related work, the motivation for this thesis, and the position of this thesis in the academic context. While the results are discussed and presented in Chapter 6, Chapter 5 gives a detailed description of how these results came about. The thesis is concluded with Chapter 7 where an outlook for potential future research is presented as well.

2 Basics

2.1 Sampled Data, Grids and (Uncertain) Fields

The goal of this section is to provide insight into the structure of scientific (sampled) data, grids, interpolation, and (uncertain) fields. The first parts of this section are based on the textbook by Telea [Tel14], please refer to it for a more detailed introduction.

2.1.1 Sampled Data, Grids and Interpolation

In general, data can be classified into two categories: intrinsically continuous or intrinsically discrete data. The latter refers to data such as websites, texts, source code, images, or any other type of record. The first, on the other hand, usually comes from nature and is measured in physical units such as kg, $\frac{m}{s}$ or something similar. Continuous data conform to the Cauchy-Criterion (also called the $\epsilon - \delta$-Criterion), which states that a function $f(x)$ is uniformly continuous if, for any small amount ϵ you choose, you can find a small distance δ such that whenever two points are within δ of each other, their function values are within ϵ of each other (see [Tel14] for the mathematical definition). Continuous data can be mathematically represented as a function in the form of:

$$f : D \to C \tag{2.1}$$

With $D \subset \mathbb{R}^d$ and $C \subset \mathbb{R}^c$. In this case, the function is called a d-dimensional, c-valued function, which means it maps from its original function domain D to values in the codomain C. Functions with $c = 1$ are called scalar fields, which assign every position in the function domain a single scalar attribute $x \in \mathbb{R}$. Vector fields ($c = 2$ or $c = 3$) on the other hand, assign to every position a vector in the form of $(v_1, v_2) \in \mathbb{R}^2$ or $(v_1, v_2, v_3) \in \mathbb{R}^3$, which can (but does not have to) depend on the original function domain. There are also fields related to tensor fields, but they are beyond the scope of this thesis.

Although the real physical field may be continuous, its computational representation is nearly always discrete. The reason for this is that it is a) hard to achieve continuous data and b) many mathematical operations (e.g., filtering, denoising, rendering) are hard

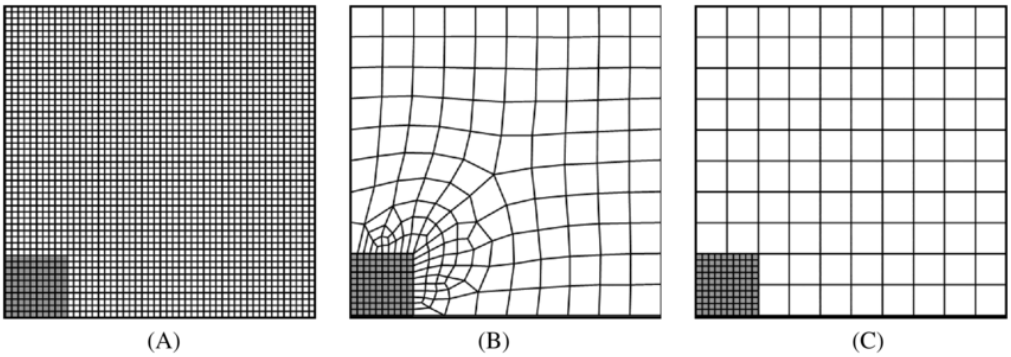


Figure 2.1: Different types of Grids: A) Uniform Grid, B) unstructured grid, and C) no-conforming grid from [KBR22]

to perform on continuous data. According to Telea [Tel14], this discrete representation is called *sampled data* and could come from, e.g., measurements or computer simulations. The sampled data can then be used to reconstruct the original continuous dataset using interpolation. Therefore, when a field is mentioned in this thesis, it usually refers to an approximation of a continuous field, such as in Equation 2.1.

Interpolation usually uses the structure of data, which is mostly called a *grid* (or mesh). A grid is a subdivision of the original function domain D into a non-overlapping collection of cells, which in turn are spanned by vertices, which are the sample points of the discretization of the continuous field. There are multiple ways of defining grids (e.g., rectilinear, structured, unstructured; see Figure 2.1) and cells (examples for 2D: line, triangle, quad, hexahedron), but for the sake of brevity this section only introduces the grid applied by the dataset used for this thesis: the uniform grid.

A uniform grid is an axis-aligned box that spans the original function domain D. The extent of the box can be described as a list of d pairs:

$$((m_1, M_1), \dots, (m_d, M_d)), (m_i, M_i) \in \mathbb{R}^2, m_i < M_i \tag{2.2}$$

(m_i, M_i) make up the lower and upper limit of the extent in each axis direction. The sample points are then uniformly distributed along the axis with a given distance δ_i depending on the axis and all the sample points p_i can be described as follows.

$$p_i = (m_1 + n_i\delta_i, \dots, m_d + n_d\delta_d), n_1, \dots, n_d \in \mathbb{N} \tag{2.3}$$

Therefore, every sample point can be described by its integer coordinates $n_1, \dots, n_d$. The number of sample points on the axis i is then $N_i = 1 + (M_i - m_i)/\delta_i$, and the set $(N_1, \dots, N_d)$ is often called the *shape* of the uniform grid. The benefits of using uniform grids are the

very low storage requirements ($3d$ floating point numbers, regardless of their size) and its simple implementation. Drawbacks are mainly that uniform grids do not represent all use-cases well or require an unnecessary high density to do so.

Interpolation is the process of reconstructing the continuous data f (Equation 2.1) from the points sampled p_i and associated values f_i. In general, there are multiple ways of interpolating, for example using the *nearest-neighbor interpolation*, assigning each point the value of the closest cell center. While this is computationally efficient, it is also a staircase-like, discontinuous approximation of the original data. A continuous approach is linear interpolation, which interpolates the value of a point $x \in D$ based on the surrounding cell linearly. But since interpolation is handled at the very last visualization step (and handled by libraries), the full mathematical description (and other interpolation ideas) are out of the scope of this thesis (but are detailed in the textbook by Telea). [Tel14]

2.1.2 Map Projections

Regarding the original function dimension d of the field f described in Equation 2.1, there is a subtle but important distinction: the *geometrical dimension* versus the *topological dimension*. The geometrical dimension refers to the dimension of space D is embedded in ($\mathbb{R}^d$), while the topological dimension refers to the actual function domain D itself, which is $s \leq d$ [Tel14]. The difference is best illustrated with the example of the application in this thesis: simulating Earth's surface. Since the earth is a three-dimensional object and the earth surface is (approximately) a sphere surface, the geometrical dimension of such datasets is $d = 3$. But since the earth's surface is a (curved) plane, the topological dimension of such datasets is $s = 2$, which is also the reason why it is sufficient to access any point on Earth with two coordinates: *latitude* (lat), referring to the degrees on the north-south axis, and *longitude* (lon), referring to degrees on the east-west axis. Typically, since the topological dimension is the most relevant one, it is referred to as the *dataset dimension*, while the geometrical dimension is assumed to be three. [Tel14]

Unfortunately, this requires mapping a 2D paper plane to a 2D sphere surface, which is impossible (without distortions) [Vie24]. Therefore, numerous different map projections were invented, which all have a different kind of distortion in different geographical places. Figure 2.2 shows two distinct projections, both with their own advantages and disadvantages. The orange circles are Tissot's indicatrices, which show the distortion introduced by the projection. An equal-area projection (Figure 2.2, right) preserves the area of the indicatrices and changes the shape. The opposite applies to conformal projections like the Mercator projection (Figure 2.2, left), where shape is preserved, but size is distorted [Gha16]. As pointed out by Vietinghoff [Vie24], a different map projection may fit the area

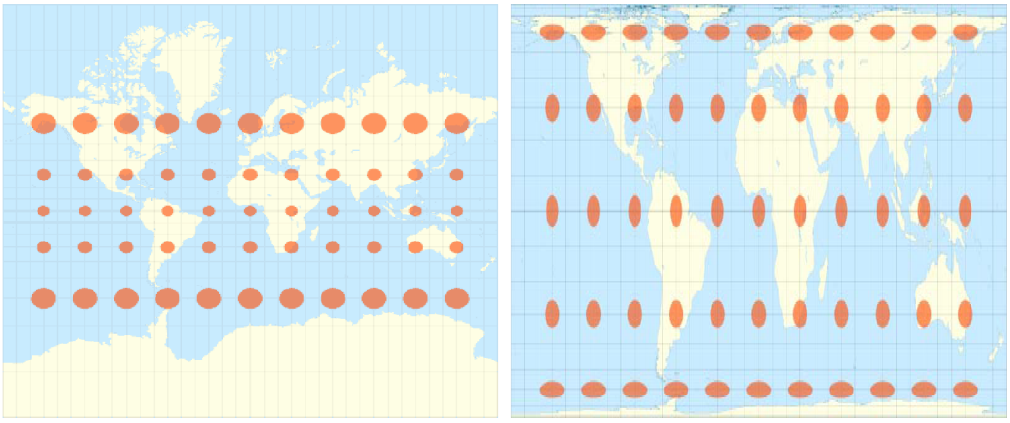


Figure 2.2: Two different map projections with different kinds of warping, depending on the map projection. The amount of warping is indicated by Tissot's indicatrices (orange circles). Left: Mercator Projection, Right: Lambert Equal-Area Projection [Gha16]

of interest better, the Mercator projection in Figure 2.2 (left) is still chosen for this thesis due to limitations in the map projection library used (see Section 5.3). This distortion also plays a role in calculations with uniform lon/lat grids, since coordinates in the far north (and south) are vastly overrepresented, because they refer to a far smaller portion of the Earth's surface than data near the equator (see Section 2.2 for geographical weighting).

2.1.3 Uncertain Fields

Let's assume $s : D \to \mathbb{R}$ is a scalar field, associating a value with every point in D. As mentioned in the previous section, such fields are usually approximated using n sample points in a grid $(x_1, \dots, x_n) \in D^n$ and their associated scalar values, represented in a vector $(s_1, \dots, s_n) \in \mathbb{R}^n$ (s_i is the scalar value at the grid point x_i). While a scalar field (or more specifically a deterministic one) has fixed values s_i, an uncertain scalar does not have explicitly defined values but instead adheres to an unknown Probability Density Function (PDF) [Vie24], as depicted in Figure 2.3.

For similar reasons why continuous data are available mostly in discrete form, in practice the PDF associated with every sample point is rarely given as the actual function, but rather as sampled values. In the case of GCMs, these samples are the result of multiple *realizations* $\omega \in \Omega$ (Ω being the sample space, which represents the different samples of the PDF), which associates every realization with one version of the field: $F : \Omega \to \mathbb{R}^n$. With regard to simulations, one realization ω is associated with a certain set of parameters for a mathematical model. Again, the complete set of Ω cannot be realized, so a specific number

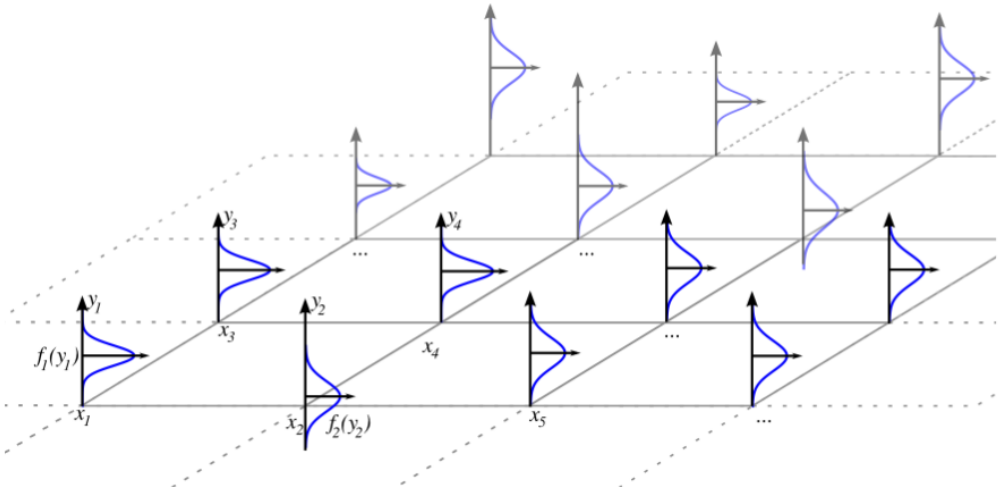


Figure 2.3: Illustration of an uncertain field with a Gaussian PDF at every sample point (from [Pöt15])

m of realizations $Z = (\omega_1, \dots, \omega_m) \in \Omega^m$ is used to approximate the PDFs associated with each point in the grid. Connecting this with the terms mentioned in Section 1.2.3: The set of fields $F(Z)$ is called *ensemble*, while each field $F(\omega_i), i \in 1, \dots, m$ is the ith *member* (or *realization*) of said ensemble. [Vie24]

2.2 Empirical Orthogonal Functions

2.2.1 Overview

Empirical Orthogonal Functions (EOF) analysis, also known as geographically weighted Principal Component Analysis (PCA) or Proper Orthogonal Decomposition [Vie24], was first introduced to the field of fluid dynamics and turbulence in the late 1960s [Wei19]. The goal was to reduce the turbulent flow to a limit number of deterministic functions to explain the structure of the flow by showing coherent structures that are otherwise difficult to find and define [Wei19]. Currently, it "is among the most widely and extensively used methods in atmospheric science" [HJS07]. One of its goals is to reduce the usually very high dimensionality of atmospheric data [HJS07], and it can also be used to link certain modes/patterns to the physics/dynamics of the analyzed system [DL02]. EOFs are a statistical procedure to decompose spatio-temporal data into two components: On the one hand orthogonal spatial patterns, on the other hand corresponding uncorrelated temporal coefficients, representing the activity of their corresponding pattern in certain time steps

[HJS07; Vie24]. The naming of the components is far from consistent: Spatial patterns are also called spatial modes, PC loadings, EOFs or even sometimes PCs, while temporal coefficients are also named principal components (PCs), EOF amplitudes or EOF (expansion) coefficients [HJS07]. So, as a formula, a spatio-temporal field $X(t, s)$ (e.g., a sea level pressure field over time mentioned in Section 1.2.2) can be described as

$$X(t, s) = \sum_{k=1}^{M} c_k(t) u_k(s) \tag{2.4}$$

with M being the number of modes/patterns and c_k the kth temporal coefficients and u_k the kth spatial pattern [HJS07]. In this work, the spatial component $u_k(s)$ is generally called the spatial pattern, while the temporal component $c_k(t)$ is called the temporal pattern or EOF coefficients. The name EOF or mode usually refers to the combination of spatial and temporal patterns.

This decomposition could be achieved in several ways, but EOF finds new sets of variables ($c_k(t)$ and $u_k(s)$ from Equation 2.4) that each capture the maximum possible amount of variance/variability of the original dataset. So, the first of the M modes captures the most variance, the second one the second most, and so on. To give a practical example: The patterns shown in Figure 1.5 show the first two spatial patterns of PSL data, the NAO and EAP (in the formula those would be $u_1(s)$ and $u_2(s)$). The temporal coefficient is basically a floating point value for each time step, indicating how active the pattern is ($c_1(t)$ and $c_2(t)$ for NAO and EAP, respectively). This value is shown for all winter seasons in Figure 1.4, second row. A positive value indicates that the PSL is dominated by the pressure low in the north and pressure high in the south, a negative value the opposite. The underlying can then be (approximately) reconstructed using $x \leq M$ EOFs (e.g., it could be approximated quite well by only using the dominant modes of NAO and EAP).

2.2.2 Mathematical Derivation and Computation of EOFs

The goal of this Section is to give an overview of the mathematical origins of EOFs based on the work of Hannachi et al. [HJS07] as well as their actual practical computation. For a more in-depth history and derivation, refer to [HJS07] and their references, while Weiss [Wei19] gives a great hands-on tutorial on POD/EOFs and their interpretation and computation.

The starting point of EOFs is usually a spatio-temporal field $X(t, s)$ defined on a grid G over n time steps, for example, the precipitation analyzed in this thesis. The value at each point on the grid at geographical location s_j and time t_i is given as x_{ij}, with $i = 1, ..., n$ and $j = 1, ..., p$. The first step is to remove the climatology of the dataset to turn it into anomaly

maps. The climatology is defined as the temporal mean $\bar{x}$ of the analyzed spatio-temporal field, so

$$\bar{x_i} = \frac{1}{n} \sum_{k=1}^{n} x_{ki} \tag{2.5}$$

$$\bar{x} = (\bar{x}_1, ..., \bar{x}_p)^T. \tag{2.6}$$

The values of the anomaly map x'_{ij} at each point in the grid are given as the deviation of X from its climatology:

$$x'_{ij} = x_{ij} - \bar{x}_j \tag{2.7}$$

And so the final anomaly map X' is:

$$X' = \begin{pmatrix} x'_{11} & x'_{12} & \cdots & x'_{1j} \\ x'_{21} & x'_{22} & \cdots & x'_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x'_{i1} & x'_{i2} & \cdots & x'_{ij} \end{pmatrix} \tag{2.8}$$

The first usual step in generating EOFs is the covariance matrix defined by

$$S = \frac{1}{n} X'^T X' \tag{2.9}$$

The covariance matrix with the values s_{ab} with $a, b = 1, \ldots, p$ contains the covariance of any grid point with any other grid point over time. To find EOFs means determining a unit length direction $u = (u_1, \cdots, u_p)$ that explains the most variability. The problem is therefore equivalent to the solution to the eigenvalue problem, which is to find all the eigenvectors (≡ EOFs) and their eigenvalue. Being an eigenvector means that the vector u multiplied by the covariance matrix S is equivalent to the multiplication with a scalar λ^2 (the eigenvalue):

$$Su = \lambda^2 u \tag{2.10}$$

To find the kth EOF of a covariance matrix, the eigenvectors u are sorted by the (largest first) value of their corresponding eigenvalue λ^2. The primary (or dominant) EOF is the first in this order, the secondary EOF the second and so on. The variance v_k of the original dataset associated with the kth EOF can then be calculated with:

$$v_k = \frac{\lambda_k^2}{\sum_{i=1}^{p} \lambda_i^2} \tag{2.11}$$

The temporal coefficients can then in turn be calculated by projecting the eigenvectors u_k on the original anomaly map X' with:

$$a_k = X' u_k \tag{2.12}$$

Together, they fulfill the requirements of the decomposition in Equation 2.4. Note here that the solutions being eigenvectors means that the multiplication by any scalar α (i.e., αu_k and $\alpha^{-1} a_k$) is also a valid solution to the problem. This indeterminacy leaves room for choosing scale and direction in a useful way (see Section 5.2 for a practical implementation) [Vie24].

2.2.3 Calculation and Application to the geographical Domain

Since geographical data are usually given on a regular 2D grid, which depicts the Earth's surface, the influence of grid point density (the same degree resolution is far more sparse in equatorial regions than in the Arctic) needs to be corrected with geographical weights. Those can be approximated by the square root of the cosine of the respective latitude [Han; Vie24] with a similar diagonal matrix as depicted in [Han]:

$$W = \begin{pmatrix} \cos(\theta_1) & 0 & \cdots & 0 \\ 0 & \cos(\theta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cos(\theta_p) \end{pmatrix} \tag{2.13}$$

Linear Algebra provides a tool called *Singular Value Decomposition* (Singular Value Decomposition (SVD)), which decomposes any matrix X into three components:

$$X = L \Lambda R^T \tag{2.14}$$

L contains the left singular vectors, R the right singular vectors, and Λ a diagonal matrix containing the singular values λ_k (as used in Equation 2.11 above).

Now, all of the above is used to calculate the EOFs of geographical data by applying SVD to the matrix (as in Vietinghoff [Vie24]):

$$\tilde{X} = \frac{1}{\sqrt{n-1}} W^{\frac{1}{2}} X' \tag{2.15}$$

When using SVD of $\tilde{X}$ with time as the first dimension (as defined at the beginning of Section 2.2.2), the columns of R^T are the EOFs (so $u_k(s)$ of Equation 2.4) and the columns of L multiplied with $\sqrt{n-1}$ are the principal components or EOF coefficients ($c_k(t)$ in Equation 2.4). As explained above, the resulting EOFs can be scaled, which is explained in detail in Section 5.2.

3 Dataset: Max Planck Institute Grand Ensemble CMIP6

3.1 Overview

The dataset chosen for this project is the MPI GE CMIP6, presented by [Olo+23]. It is a single-model initial-condition large ensemble (in short: SMILE) consisting of multiple, coupled models: ECHAM6[1] for the atmosphere directly coupled to JSBACH[1] for land and MPIOM[1] for sea and sea-ice. The models are coupled once a day, which means that the simulation results of the different models serve as input for the other models. As an ensemble simulation, it consists of multiple members, which are different variants of the simulation with the same forcings (such as GHGs) but different initial conditions. To generate the initial conditions, the historical simulations are split from 1000-year quasi-stationary preindustrial control simulation circa 25 years apart for each member, and the results of them in the year 2015 serve as the initial state for each corresponding member in the future scenarios. [Olo+23]

Differences from its predecessor MPI GE [Mah+19] (which was used in the work of Vietinghoff [Vie24]) include improved time resolution (from monthly means up to 3 or 6 hour intervals) and the updated CMIP6 forcings and future scenarios (see Section 3.2). Since MPI GE CMIP6 follows the CMIP6 protocol (see Section 1.2 and [Eyr+16]), it implements the DECK core with (among others) a quasi-stationary preindustrial control simulation and historical simulations. Furthermore, it also uses the forcings defined by CMIP6 (like volcanic eruptions, solar circle, GHGs etc.) for the historical and future simulations (see Section 3.2).

[1] Labels of the different models

3.2 ScenarioMIP: Future Scenarios and Shared Socioeconomic Pathways

Since the goal of this thesis is to evaluate the prospects of climate change, simulations of the future are necessary. CMIP (Phase 3) introduced a project of future climate scenarios (ScenarioMIP) in the 2000s, which defines and simulates the developments of different anthropogenic drivers of climate change [ONe+16]. They play an important role in climate research and have since been the source of many figures and assessments in IPCC reports [TBL20]. The different scenarios can be used to assess "possible changes in the climate system, impacts on society and ecosystems, and the effectiveness of response options such as adaptation and mitigation under a wide range of future outcomes" [ONe+16]. The differences between the scenarios are the forcings introduced by multiple variables, including change in land use, climate change mitigation policies, energy use, population, economic growth, and emissions [Ria+17]. For CMIP6, they extended the old model of Representative Concentration Pathwayss (RCPs), a predefined ERF reached in 2100, by adding so-called Shared Socioeconomic Pathwayss (SSPs). These SSPs add socioeconomic reasons for the assumed changes in land use and emissions.

SSPs are derived from five broad abstract narratives, which are then quantified in different ways. So, for example, the narrative for SSP1 is: "Sustainability – Taking the Green Road (Low challenges to mitigation and adaptation) The world shifts gradually, but pervasively, toward a more sustainable path, emphasizing more inclusive development that respects perceived environmental boundaries. Management of the global commons slowly improves, educational and health investments accelerate the demographic transition, and the emphasis on economic growth shifts toward a broader emphasis on human well-being. Driven by an increasing commitment to achieving development goals, inequality is reduced both within and between countries. Consumption is oriented toward low material growth and lower resource and energy intensity." [Ria+17]

In contrast to this, the narrative for SSP5 is: "Fossil-fueled Development – Taking the Highway (High challenges to mitigation, low challenges to adaptation) This world places increasing faith in competitive markets, innovation and participatory societies to produce rapid technological progress and development of human capital as the path to sustainable development. Global markets are increasingly integrated. There are also strong investments in health, education, and institutions to enhance human and social capital. At the same time, the push for economic and social development is coupled with the exploitation of abundant fossil fuel resources and the adoption of resource and energy intensive lifestyles around the world. All these factors lead to rapid growth of the global economy,

while global population peaks and declines in the 21st century. Local environmental problems like air pollution are successfully managed. There is faith in the ability to effectively manage social and ecological systems, including by geo-engineering if necessary." [Ria+17]

The narratives of SSP2 to SSP4 lie somewhere between (see [Ria+17]). These narratives are then quantified in multiple dimensions (resource availability, technical development, lifestyle changes, population, economic activity, etc.). These quantifications then serve as input for a variety of integrated assessment models (IAMs), which turn them into the actual forcings needed (e.g., land and energy use, emissions) [Ria+17].

In actual scenarios, these SSPs are combined with additional radiative forcing (RCP, the earlier version of the scenarios in CMIP5), resulting in a matrix which can be seen in Figure 3.1. So, a scenario using the narrative of SSP5 and an ERF of 8.5 Wm^{-2} in 2100 is now called SSP585. Although there are now 35 possible scenarios, O'Neill et al. defined two different tiers of scenarios ranked according to their importance. Figure 3.1 lists Tier 1, which are scenarios mostly comparable to the old RCP scenarios. These scenarios are available in the MPI GE CMIP6, amongst some of Tier 2. [BK; ONe+16; Ria+17]

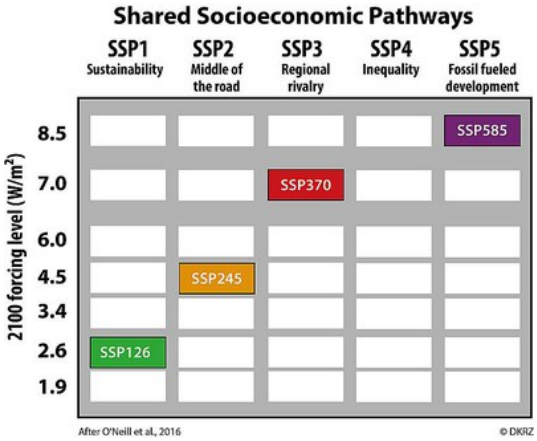


Figure 3.1: Combinations of SSPs and RCPs leading to scenarios comparable to the old RCPs. The vertical axis describes the old RCP variant of a forcing level in 2100, while the horizontal axis are the SSPs 1 to 5. In combination, they define the new scenarios. [BK]

3.3 Dataset description

3.3.1 Resolutions and Dimensions

In terms of spatial resolution, MPI GE CMIP6 comes in three variants: The low resolution variant MPI-ESM1.2-LR with a horizontal resolution of roughly 1.8° longitude/latitude res-

olution in the atmospheric part and 0.4° lon/lat for the ocean, the high resolution variant MPI-ESM1.2-HR with a horizontal resolution of 1.0°/0.4° for atmosphere/ocean and the extreme high resolution MPI-ESM1.2-XR with 0.5°/0.4° for atmosphere/ocean. Each variant has a vertical resolution of 47 levels for the atmosphere and 40 levels for the ocean. With increasing spatial resolution comes decreased availability of other variables such as simulation members, covered time period, and implemented scenarios. Although [Olo+23] reports 30 members for each simulation (for the LR variant), in the actual dataset available for this work, 50 members were simulated.

In terms of time resolution, MPI GE CMIP6 provides very few, limited variables in 3 hour intervals and most variables in a 6 hour interval. A complete list of the variables can be seen in [Olo+23, Table 3], the variables necessary for this thesis are listed in Table 3.1.

Table 3.1: Variables necessary for this thesis, derived from [Olo+23]

Name	Parameter Long Name	Unit	Vertical Levels
hus	Specific Humidity	1	47
ua	Eastward (Zonal) Wind	ms^{-1}	47
va	Westward (Meridional) Wind	ms^{-1}	47
ps	Surface Air Pressure	Pa	1
pr	Precipitation	$kg\,m^{-2}s^{-1}$	1
psl	Sea Level Pressure	Pa	1

3.3.2 Vertical Hybrid Sigma Pressure Layers

Regarding vertical levels, all variables were not available in fixed pressure layers but in the so-called *hybrid sigma pressure coordinates*. In comparison to fixed pressure layers (such as 1000 *hPa*, 750 *hPa*...), hybrid sigma pressure coordinates follow the terrain (mountains, valleys, etc.). Essentially, sigma vertical levels are given as fractions of the surface pressure P_S at any point, following the equations in [Eck09]:

$$\sigma = h(p, P_S) = \frac{p - P_{top}}{P_S - P_{top}} \tag{3.1}$$

Here, $p \in [P_S, P_{top}]$ is a pressure level. It was proposed that instead of providing pressure levels at pure fractional levels, it would be better to smoothly converge from terrain following fractions (sigma levels) at lower (i.e., near the earth surface) levels to isobaric (i.e., same pressure) levels in higher altitudes. This approach has both numerical and practical

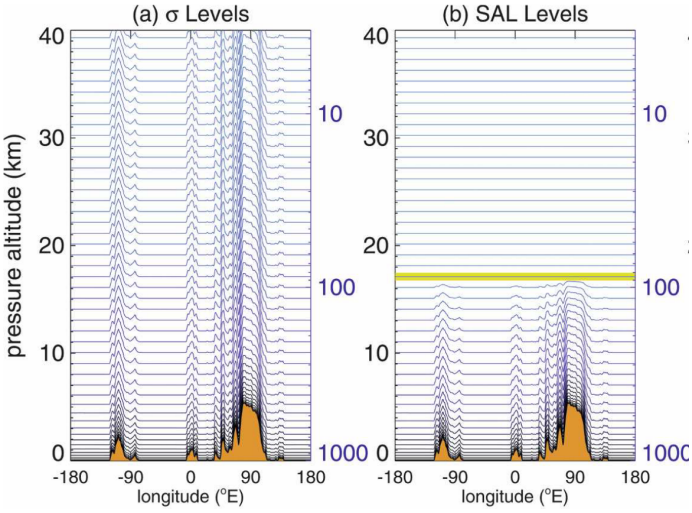


Figure 3.2: Examples of (hybrid) sigma pressure layers. a) shows sigma layers like in Equation 3.1, while b) shows a hybrid approach in the form of Equation 3.2 [Eck09]

advantages. There are also alternatives for $h(p, P_S)$, but a final form[2] is used to calculate pressure levels at any discrete sigma level $\tilde{\eta}$:

$$p(\tilde{\eta}, P_S) = A(\tilde{\eta}) + B(\tilde{\eta})(P_S - P_{top}) \tag{3.2}$$

$A(\tilde{\eta})$ is a vertical shift, which is close to zero at low levels, and $B(\tilde{\eta})$ is a fraction of the pressure range $(P_S - P_{top})$, which is close to zero at high levels.

Using Equation 3.2 results in levels of equal pressure thickness, which merge to isobaric layers at higher altitudes. Olonscheck et al. [Olo+23] does not report their exact approach to $A(\tilde{\eta})$ and $B(\tilde{\eta})$ in MPI GE CMIP6, but every data set that uses hybrid sigma pressure levels contains variables $ap(lev)$, $b(lev)$ and the pressure level field $ps(lon, lat, time)$, from which the pressure at any point and level can be calculated with:

$$p(lev, lon, lat, time) = ap(lev) + b(lev)ps(lon, lat, time) \tag{3.3}$$

The upper pressure limit can be ignored in Equation 3.3 since the upper border is zero in MPI GE CMIP6.

[2]The equations that lead to that can be seen in [Eck09]

3.3.3 Structure of the data

The data is available in the high performance computing cluster of the DKRZ[3], the structure can be seen in Figure 3.3.

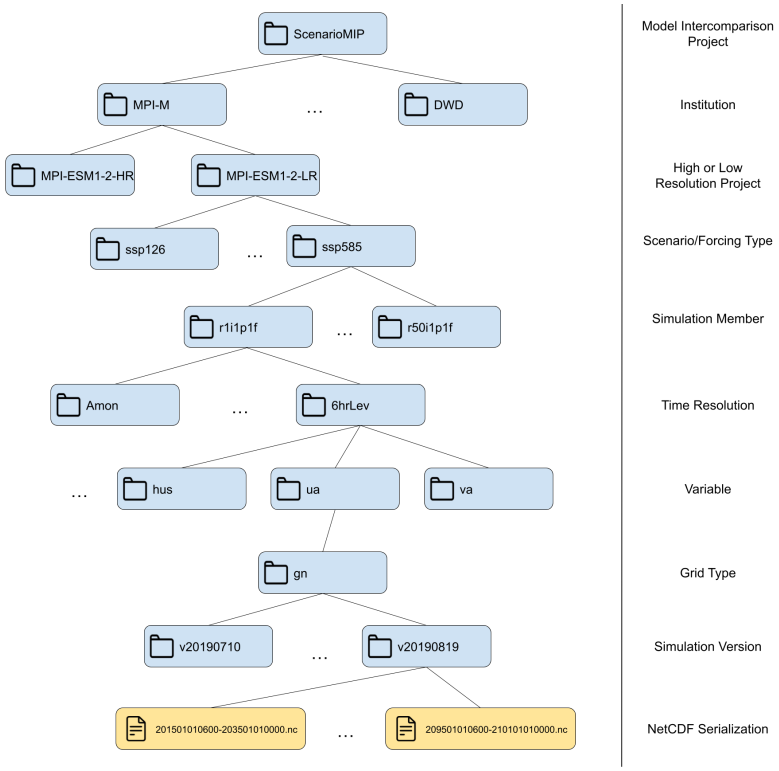


Figure 3.3: Structure the data is available on the DKRZ cluster. Example is given for ScenarioMIP, but applies as well for historical and piControl.

The root folder is a MIP, for example, ScenarioMIP or the CMIP core, followed by the institution and resolution category (see Section 3.3.1). After a hierarchy of forcing types (e.g., SSP, historical, piControl), member IDs and time resolutions, the variable (e.g., *hus* for specific humidity) can be selected. After the grid type (only *gn* is available) the version directory contains the actual data, serialized in the NetCDF4 format (which is based on HDF5 [Fol+11]) and divided into time scopes of up to 20 years. The version is named after the date of the simulation and, since a later version indicates a fix in the older version, it is generally advisable to choose the latest version for each variable.

[3]Deutsches Klimarechenzentrum (en.: German Center for Climate Calculation)

3.3.4 NetCDF Datasets

The goal of the Network Common Data Format (NetCDF) was to create a machine-independent format for representing scientific data. It consists of an abstraction for storing multidimensional data, an implementation of said abstraction with a data format, and a library supporting that data format. It is modeled for supporting scientific datasets consisting of multiple, named, multidimensional variables together with their reference grid/coordinate system and some metadata properties. Every variable consists of a type (e.g., scalars, byte arrays, characters, floating-point numbers), a shape defined by a vector of dimensions, and auxiliary properties as key-value pairs (e.g., physical unit, other names, important notes). [RD90]

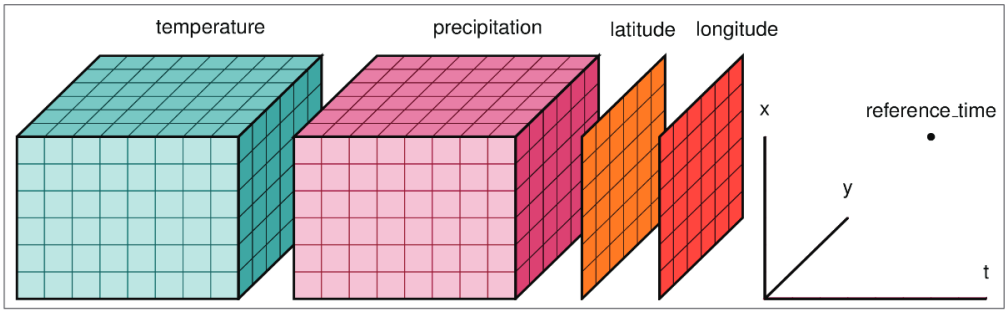


Figure 3.4: An example of a named multidimensional dataset from [HH17]: Precipitation and temperature are the variables while x, y and t are dimensions. Longitude and latitude are coordinates (also variables) defined by x and y.

Figure 3.4 shows an example of the said structure in the form of a meteorological dataset: x, y, and t are the dimensions, named integers representing the shape of variables. Precipitation and temperature are three-dimensional variables, while longitude and latitude are coordinates of the grid, giving a reference for the location of the grid. The dimensions of the dataset are x, y and t, giving both variables the shape (x, y, t). [HH17]

4 Related Work

This section outlines the current state-of-the-art in the main parts of this thesis explained in Section 1.3: Quantifying Moisture (Transport), extracting spatio-temporal patterns, tracking their change over time, and visualizing the uncertain results in the end.

4.1 Motivation

As explained in Chapter 1, the approach of this thesis is motivated by the approach of Vietinghoff et al. in [Vie+21a] and the affiliated dissertation [Vie24], which tackles the issue of detecting critical points in unstable scalar fields. Here, [Vie+21a] analyzes the MPI GE [Mah+19] from the fifth phase of CMIP, an ensemble simulation with 50 members. The goal was to find the probable high / low pressure centers in the NAO pattern (see Section 1.2) and to track their shift over time. They used a sliding window approach, computing the dominant pattern (see Section 2.2) for each window and member, and determining the likely areas of critical points by merging the results of different members per time step. The centers of mass of these critical areas are then tracked over time to visualize the shift of pressure highs and lows. The results show that the patterns change and that this change is more pronounced if the climate change is more pronounced. Also, there is no significant change if the climate remains stable.

4.2 Moisture Transport

In order to computationally extract any spatio-temporal pattern of moisture (transport), the moisture transport must be quantified in some way. The variable of MPI GE CMIP6 used for this task is the *specific humidity*, which has no unit and is a floating value between 0.0 and 1.0. Specific humidity denotes the percentage of water in the air at a given point on the grid. The vast majority of literature on moisture transport uses some form of vertically integrated humidity, and the variants are explained in the following section. A popular use of these quantifications is to find a filamentary weather structure called

"Atmospheric Rivers"[1], a prominent way of transporting water vapor in the extratropical regions [Gim+14].

The most straightforward approach to quantifying moisture is **Vertically Integrated Water Vapor (IWV)** [Bao+06; Gim+14; Nei+08; SE90; WBR18; Zha+21](also known as perceptible water [WBR18]), which is the vertical integral of the specific humidity q over pressure levels p from the surface of the earth P_S to some upper limit in the atmosphere:

$$IWVs = \frac{1}{g}\int_0^{P_s} q \ \mathrm{d}p \tag{4.1}$$

Similarly to Equation 4.1, Zhu and Newell proposed in [ZN98] to use **Integrated Water Vapor Transport (IVT)** for the detection of atmospheric rivers. It is calculated by vertically integrating the zonal (along latitude lines) and meridional (along longitude lines) fluxes over the different pressure levels. It became a popular metric for finding atmospheric rivers [Gim+14], sometimes alongside IWV [EBM16]. IVT has the unit $\frac{kg}{ms}$ and is usually defined with

$$\overrightarrow{IVT} = \frac{1}{g}\int_0^{P_s} q \begin{pmatrix} u \\ v \end{pmatrix} \mathrm{d}p \tag{4.2}$$

or in a mathematically equivalent form [FSZ03]. Here, u and v represent the zonal and meridional components of the horizontal wind vector. An equivalent way is to calculate the zonal and meridional components separately with

$$IVT_z = \frac{1}{g}\int_0^{P_s} q\, u \ \mathrm{d}p \tag{4.3}$$

$$IVT_m = \frac{1}{g}\int_0^{P_s} q\, v \ \mathrm{d}p \tag{4.4}$$

While Equation 4.2 yields a vector field, the Euclidean norm of the vector field

$$\|IVT\| = \sqrt{(IVT_z)^2 + (IVT_m)^2} \tag{4.5}$$

is also a popular choice in detecting atmospheric rivers [Lan+24; Ram+16; Sou+20] and other use cases, such as researching patterns of moisture transport [Aya+22; KA15; Zou+20; ZY05] (see the next Section for details).

[1] Earlier or alternative name: "Tropospheric Rivers"

The IVT is also part of the atmospheric moisture budget [Yan+22] (and similar in [Sea+20]) given by

$$\frac{1}{g}\frac{\delta}{\delta t}\int_0^{P_s} q \, \mathrm{d}p = -\nabla \cdot \frac{1}{g}\int_0^{P_s} q \begin{pmatrix} u \\ v \end{pmatrix} \mathrm{d}p + E - P \tag{4.6}$$

With E being the total evaporation and P the precipitation. Yang et al. showed in their report [Yan+22] the directions of moisture flux and its evolution in the last three decades. The analysis was done for all continental borders based on the ERA5 reanalysis. The metrics used for this analysis were mainly the evaporation E, precipitation P and the convergence of moisture transport $VIMC = \frac{1}{g}\int_0^{P_s} \nabla \cdot q \begin{pmatrix} u \\ v \end{pmatrix} \mathrm{d}\, p$ from Equation 4.6.

Although the integration in the previous equations integrates from the surface to the outer border of the atmosphere ($0hPa$), it is quite common to integrate up to the limit of $300hPa$ [Aya+22; Gui+18; KA15; ZN98], since the amount of moisture in the regions from $300hPa$ to $0hPa$ is negligible and amounts in total to about 2-3 cm/year in terms of freshwater flux [ZY05].

There are other notable algorithms, namely stable oxygen isotope investigation [Ma+18] and Langragian backward trajectories [Zha+21], but both look for the origin of water vapor instead of its destination and are therefore beyond the scope of this thesis.

4.3 Pattern analysis regarding IVT

Although there are many areas of interest for the application of EOF, this section will give an overview of what kind of pattern analysis has been performed in relation to moisture transport data. This is not limited to patterns of moisture (transport), but also to calculating patterns of other variables and linking/comparing those to the moisture (transport). The procedure is quite similar in most related work:

1. Generate the EOFs and visualize the spatial alongside the corresponding temporal pattern for an overview.
2. Use other variables, e.g., precipitation, PSL (representing an oscillation like the NAO), occurrence of atmospheric rivers, or Sea Surface Temperature (SST), to interpret the patterns. For this process methods like linear regression or (cross)correlation are used to explore the relationships between different variables. Those methods are usually applied on the temporal patterns of the analyzed mode and the actual data of other variables. Other variables typically include indexes of oscillations like ENSO [Aya+22; KA15] or the raw data of precipitation.
3. Visualize the results using maps and diagrams.

An overview of datasets, time scopes, and other metadata from similar work is given in Table 4.1.

Published in 1982, Salstein et al. [SRP83] provided the first example of calculating EOF of IVT. Based on data from 91 weather stations, they computed the IVT of the entire northern hemisphere. Statistical significance was determined using a Monte Carlo testing method. The EOFs were computed on the IWVs, the zonal and meridional IVT fields respectively, but they also evaluated an approach of combining both IVT components in one data vector. They particularly reported the significance of the primary mode of IWVs, encoding nearly half (44 %) of the variance of the data.

Although most of the related work found uses EOF analysis, Teale and Robinson employ an approach using Self Organizing Maps (SOMs) to detect moisture transport patterns in the eastern United States. SOMs are a machine learning approach to reduce data dimensionality, producing a 2D map of higher-dimensional data. Although they acknowledge the efficiency of EOF in extracting dominant patterns, they emphasize the problem of required orthogonality of modes, which is not given for SOMs. The results show that fluxes with the highest moisture content occur less frequently than those with less moisture. However, despite the higher moisture content, fluxes with lower moisture transport dominate the water vapor movement because of their prevalence. Many of these fluxes meet typical criteria for atmospheric rivers, with varying trajectories and sources suggesting various mechanisms of formation. The temporal variability in monthly flux frequencies correlates with regional precipitation patterns, indicating that this approach is a valuable framework for studying precipitation changes. [TR20]

Ayantobo et al. analyzed the main six modes of EOF in China, which were divided into seven different regions for comparison. While the variances of IVT in eastern to southern China were quite high, the variances in northern China were quite low. By comparing the temporal patterns of the primary mode of EOF with that of ENSO, it was shown that these patterns were related. Cross-wavelet coherence[2] showed that the IVT and ENSO time series were coherent, implying that increased IVT was generally associated with increased ENSO activity. [Aya+22]

[WBR18] compares the patterns of perceived water (IWVs) in Europe for different seasons/months for the last ≈ 50 years. Similarly to [Aya+22], Europe was grouped into different regions with different moisture conditions. This revealed significantly different moisture patterns for the regions, for example, the northern continental vs. the northern Atlantic. The results confirmed the important expected role of atmospheric circulation (rep-

[2] Also known as time-variation Fourier analysis, this approach decomposes signals into the time-frequency domain to analyze where in time (x-axis) the signals are related at what frequency (y-axis). [Aya+22]

resented in this study by EOF patterns of PSL and the advection direction of air masses) for moisture in winter by measuring correlation, while the relationships were much weaker for transitional months like April or October. The lack of correlation between atmospheric circulation and moisture patterns in summer was also striking. [WBR18]

Fernández et al. analyzed the precipitation modes in the Mediterranean Sea and related them to the transport of moisture in the same area. The purpose of this analysis was to contribute to the understanding of the precipitation reduction that occurred in the area, as well as the low-frequency variability of precipitation that led to multi-year droughts. They used several methods to validate their data: The precipitation data and the wind/moisture data for IVT were validated with data from actual weather stations. The stability of the eigenvectors was tested with a Monte Carlo simulation that compared the variability of the actual data with random test data, while the degeneracy of the EOF modes was tested using the method of North et al. [Nor+82]. The results of the analysis identify the interpretation of the three main precipitation modes: The first mode (22 % variance) appears to be related to the NAO, Atlantic storm tracks and associated moisture transports, while the second mode (16 %) represents the internal redistribution of moisture in the Mediterranean basin between the eastern and western parts. The third mode (11 %) explains the increased precipitation in the northern part of the domain. In addition, moisture transport during the positive and negative phases of the leading mode showed an increased inflow of moisture from the west. [FSZ03]

Similarly to the work of Fernández et al. [FSZ03], Zhou and Yu analyzed the anomalous summer rainfall patterns over China and linked them to water vapor transport. They confirmed their results using a second dataset for IVT calculation. Their work showed that the primary mode of anomalous rainfall is associated with heavier rainfall in the Yangtze river region, while the same applies to the second mode and the Huaihe river. Connecting these patterns with moisture transport, they identified the different ways in which these heavier rain areas are caused by certain convergences of water vapor transports. Furthermore, they compared the supply of anomalous rainfall patterns with that of normal monsoon rainfall, revealing that these differ significantly. [ZY05]

In the work of Guirguis et al. [Gui+18], the authors calculate rotated EOF of IWV data and try to analyze the relationship between the 15 most dominant modes and the occurrence of atmospheric rivers (AR) on the west coast of the United States. For this, they divided the coast into different regions and linked the activity (positive and negative) of the corresponding temporal pattern of each mode to the occurrence of atmospheric rivers. It was found that some modes seem very influential for some regions' AR activity, while others seem to play no role at all. They also identified favorable and unfavorable circulation

states (e.g., among others, a low-pressure anomaly in some regions) for the occurrence of AR [Gui+18].

Kim and Alexander showed in their analysis the connection of IVT patterns in the western US to three different ENSO events (Eastern Pacific El Niño (EPEN), Central Pacific El Niño (CPEN) and La Niña (NINA)). Although EPEN events are associated with large positive IVT anomalies from the subtropical Pacific to the northwest of the United States, CPEN events lead to enhanced moisture transport to the southern United States. During NINA events, the mean IVT anomaly is reversed compared to EPEN and CPEN. It is also shown that the IVT patterns computed for these events are significantly different from those computed for neutral years. Furthermore, the results were related to precipitation anomalies on the west coast of the USA, showing large differences (especially for the northern part of the coast) for EPEN and CPEN events. However, the authors also emphasize that while the evidence is strong, there are exceptions (e.g., one El Niño leads to a dry winter, another to the opposite) and need to be studied in more detail. [KA15]

Similarly to [Vie+21a] and the approach of this thesis, Zou et al. applied a sliding window approach to IVT patterns in the tropical Indian Ocean–western Pacific to analyze the evolution over time. For the studied period from 1961 to 2015, they studied every 20-year period with a 5-year sliding window, computing Multivariate EOFs for each window, resulting in vector fields of patterns. The results show that the two most significant modes show significant changes in the mid 80s: The primary mode is characterized by an anticyclonic pattern in the north-western Pacific, which shifts significantly to the south. An analysis of the relationship with sea surface temperature (SST) revealed that the correlation between the mode and SST rose in the mid 80s, from weakly correlated to a significant positive correlation between IVT and SST anomalies. Furthermore, the primary mode appears to be significantly regulated by ENSO. The second most significant mode is related to the variability of the tropical Indian Ocean dipole (defined by the differences in average SST) [Zou+18].

A different approach was employed by [Zou+20], evaluating the EOF patterns of vertically integrated apparent moisture sinks. Results indicate that the primary mode is a southwest-northeast oriented dipole, while the secondary mode is a southwest-northeast oriented tripole. The primary mode seems to be heavily regulated by ENSO in the previous winter season, while the second mode seems to originate from internal atmospheric variability. Based on the much higher standard deviations in ENSO years, it seems that the water vapor source and sink tend to be dominated by the primary mode in ENSO years, while the secondary mode is prevalent in non-ENSO years.

Table 4.1: Overview table of research regarding patterns with moisture transport. THe acronyms in the Studied Season column stand for the month used: JJA for the summer (June, Juli, and August) and (N)DJF for winter (November, December, January, and February). TEIOWP stands for Tropical Eastern Indian Ocean-Western Pacific, the region around Indonesia.

Release Year	Pattern extraction	Area of Interest	Timescope	Time Resolution	Studied Season	Variable used for EOF
2020 [TR20]	SOMs	eastern USA	1979 to 2017	daily	all year	$\|IVT\|$
2022 [Aya+22]	EOF	China	1979 to 2010	daily	all year	$\|IVT\|$
1982 [SRP83]	EOF	Northern hemisphere	1958 to 1973	monthly/yearly	all year	IWV, IVT_m, IVT_z, combined
2003 [FSZ03]	EOF	Mediterranean Sea	1948 to 1996	6hr	DJF	PR
2005 [ZY05]	EOF	China	1951 to 1999	monthly	JJA	PR
2018 [Gui+18]	EOF	USA (west coast)	1948 to 2017	daily	NDJF	IWV
2015 [KA15]	EOF	western USA	1979 to 2010	6hr	DJF	$\|IVT\|$ (assumed)
2018 [Zou+18]	EOF	TEIOWP	1961 to 2015	monthly	JJA	$\overrightarrow{IVT}$
2020 [Zou+20]	EOF	TEIOWP	1958 to 2018	6hr/monthly	JJA	Integrated Water Vapor Sink
2013 [Yao+13]	EOF	East Asia	1997 to 2002	-	JJA	IVT_m, IVT_z
2012 [LZ12]	EOF	East Asia	1979 to 2009	monthly	summer	$\overrightarrow{IVT}$
2018 [WBR18]	EOF	Europe	1981 to 2015	daily	all year	IWV, SLP

Although the main focus of [Yao+13] is to evaluate and compare a regional air-sea coupled model, they also performed EOF analysis on the zonal and meridional components of IVT, respectively. They used the results to evaluate the connection to SST, revealing that the results of the regional coupled model align better with the results of other data sets and reality than the regional uncoupled model.

Li and Zhou evaluated the connection of the IVT-EOF patterns to ENSO in the northwestern Asian Pacific. They used a different approach than most in applying EOF to IVT, by concatenating the meridional and zonal components in one matrix and calculating EOF on it. To confirm their results, they compared the results with another reanalysis from the same (and larger) region. Furthermore, these IVT patterns were linked to the SST. They revealed the characteristics of the two most significant modes, but most prominently they showed the quasi-4-year coupling of the two most prominent modes with ENSO [LZ12].

4.4 Uncertainty Visualization

Since the data set used (cf. Chapter 3) is an ensemble simulation consisting of 50 members, most of the figures and other visual representations in this thesis need to display the uncertainty arising from them. This section summarizes advances fitting for this topic, giving a framework of references of current possibilities of visualizing uncertainty.

Kamal et al. [Kam+21] give a recent overview over the whole topic of uncertainty visualization: From the introduction to the whole concept of uncertainty to the differentiation between different kinds of uncertainty in the visualization process. On the one hand, there is the uncertainty of the data itself, resulting from measurement errors, sensor imprecision,

or simulations. On the other hand, there is the uncertainty introduced by the visualization itself, such as interpolation between grid points. They grouped all kinds of uncertainty representation into two categories: quantification, consisting of mostly mathematical approaches of handling uncertain data, and visualization, which displays the uncertain data directly. An example for the former would be to reduce the original data using various techniques to make it possible to display the PDF in the end (e.g., Pöthkow et al. [PWH11]) or using clustering techniques to identify outliers (e.g., [BKS04]). For the latter, different types of uncertainty visualization were presented and reviewed. They come in various forms: For example, animating, like introducing vibrations to show more or less uncertain areas, like in the work of Brown [Bro04], as well as manipulating speed, blur, or blinking. Also, changing visual variables (such as color, hue, brightness), which was reviewed in the empirical study of MacEachren et al. [Mac+12]. Another popular way is to display the uncertainty with graphical techniques such as box/scatter plots and glyphs. The boxplots, for example, can be altered to show the uncertainty on the outer lines in various ways, e.g., the work of [Ben88]. Also, boxplots can also be depicted using a 2D map, where the distortion in the 3rd dimension represents the uncertainty [Kao+02]. Another way can be to use glyphs, where different properties, such as size, shape, or complexity, can also be altered to show the uncertainty in different places. Furthermore, recent advances in uncertainty visualization are given, with a special emphasis on ensemble (simulation) data, big data, and machine learning, listing the most prominent areas where the presentation of uncertainty is crucial. In the end, a framework for evaluating uncertainty visualization is presented, followed by an overview of possible future research directions [Kam+21].

The way in which animation can display uncertainty in scalar fields was shown by Coninx et al. [Con+11]. Their goal was to enrich the usual display of scalar fields with color maps with additional uncertainty information. The tool of choice here was animated Perlin noise, and the uncertainty was presented by modifying the noise mask with the uncertainty information at each point. The results were tested using a psychophysical evaluation of contrast sensitivity thresholds, evaluating the effective parameters for the proper presentation of the uncertain area. [Con+11]

Sanyal et al. proposed Noodles, a tool for displaying uncertainty in weather ensemble simulations. It employs three different ways of displaying uncertain isocontours: ribbon, glyphs, and spaghetti plots. In addition, they added tools for exploring the uncertainty of data sets, such as a color map of the entire uncertainty of the data set. Uncertainty in spaghetti plots is clear (one line per member) but gets confusing and chaotic quickly. The glyphs display the uncertainty by different sizes and can be displayed on the whole map or alongside means of isocontours. Ribbons condense the information of multiple lines by

adapting the width of the ribbon to the uncertainty of isocontours at a specific point in the grid. The resulting tool was tested by two meteorologists and they classified the results as beneficial. [San+10]

Another way of visualizing groups of isocontours is using contour boxplots proposed in [WMK13], which group isocontours similar to conventional boxplots. This means that the easiest default presentation (spaghetti plots) is replaced by popular boxplot stats: The median, the mean, the quartiles around that mean, the whole range, and the outliers (not part of the whole range). However, the implementation is not as straightforward as in conventional boxplots. To quantify the aforementioned statistics, Whitaker et al. propose a data depth-based approach, which encodes how much a particular sample is centrally located in its function (or in this case: How central is an isocontour to a whole set of isocontours). While the results look very promising, the algorithm lacks a publicly available implementation, making it difficult to use the approach.

Also notable in this section is the recent work of Lan et al. [Lan+24], which aims to display the topology and uncertainty of atmospheric rivers (ARs). The uncertainty here does not stem from ensemble simulations, but rather from the numerous ways of defining and identifying ARs. They used a set of algorithms that usually use different spatial, temporal, and IVT thresholds to identify ARs and their spatial location. The uncertainty is displayed using an implementation of contour boxplots proposed in [WMK13]. Furthermore, the topological skeleton is extracted and displayed. The uncertainty in the topological skeletons is then displayed using an algorithm of another paper, which especially emphasizes the regions of agreement and disagreement across the ensemble of ARs. Feedback from domain experts and case studies shows that both approaches complement each other and are an effective way to comparatively display atmospheric rivers. [Lan+24]

4.5 Position of this Thesis

As shown in Section 4.3, Empirical Orthogonal Functions are a relatively popular tool for analyzing spatio-temporal patterns in moisture transport. Although mostly applied to water vapor transport in South East Asia and the Chinese Sea, there is not much coverage of the European Area, and especially the larger scope of the north-east Atlantic (except for the work of Wypych et al. [WBR18]). Furthermore, there has been no EOF IVT analysis for ensemble-scale data, just for reanalysis data (see Table 4.1) or actual weather station data. Analyzing the evolution of spatio-temporal patterns is also quite underrepresented, since most approaches apply the pattern analysis on the entire available dataset (in most cases around 50 years), exceptions are the motivational work for this thesis from Vietinghoff

et al. [Vie+21a] and Zou et al. [Zou+18]. Additionally, to the authors' knowledge, no future scenario pattern analysis has been performed with IVT, especially not with data from CMIP datasets.

In terms of uncertainty visualization, most of the approaches presented could be quite useful in this thesis: The animated Perlin noise from [Con+11] could be used to show the uncertainty in the scalar fields, while the ribbons and contour boxplots can be used to represent contours in the patterns (see Section 5.3 for the decision on feature extraction). Unfortunately, none of these algorithms is available as a library, which hinders application to a great extent.

So, this thesis tries to fill the identified gap in related work in the following way: Implement a sliding window EOF analysis to study the evolution of moisture-related patterns (**M1**, similar to [Vie24; Zou+18]). The variables chosen for this are of course the IVT, but also the sea level pressure, representing the most influential oscillations of that area (NAO and EAP), and precipitation, as it is very significant for ecological and economic reasons and is one of the most popular choices of related work. The next step is to compare the relationships of patterns and variables using (cross)correlation and/or linear regression (**M2**), similar to [Yao+13; Zou+18; Zou+20]. Although it is easy to compute and visualize the comparison of EOF patterns of different variables, Dommenget and Latif [DL02] provide a good example of why EOF patterns are difficult to interpret and link to actual physical modes. They recommend using multiple evaluations and statistical tools (such as regression) to link the mathematical modes produced by EOF analysis with the actual existing physical modes in the real world. Although this has been done for the EOFs of Sea Level Pressure in the Northern Atlantic (see Section 1.2.2 and especially Figure 1.4, the NAO index calculated from weather stations is structurally very similar to the first principal component), no analysis like this exists (to the authors' best knowledge) for patterns of IVT and precipitation in Europe. So, in order to interpret the results, it is important to make sense of them by checking their relationship with other data. In the end, the results need to be visualized (**M3**), with the challenge of displaying the variability introduced by multiple members of the ensemble.

5 Methodology

This Section gives a detailed description how the analysis was performed on the basis of the MPI GE CMIP6, how the huge datasets were preprocessed, the EOF computed, and the result visualized.

5.1 Preprocessing

The purpose of this step is to prepare the datasets for pattern generation. In addition to the calculation of moisture transport, this step has the following goals: The reduction of the dataset to the area of interest (northern Atlantic and Europe) and the simplification of the directory structure shown in Figure 3.3 to make usage easier. Additionally, the 6-hourly dataset is reduced to monthly time resolution to be able to calculate the stationary and transient components of moisture transport (cf. Section 5.1.2 for an explanation).

The calculations were performed in the high performance computing (HPC) cluster[1] of the German Climate Calculations Center (DKRZ), since the MPI GE CMIP6 is located there and downloading the data would take an unfeasible amount of time. Since they are billed by the time of node usage, another goal of this step is also to minimize the hours of use of the HPC system. Due to the large size of the datasets (resulting from the high time resolution, multiple members, and scenarios), the process needs to be implemented efficiently.

5.1.1 Chosen Framework

The goal of this step is to prepare the data for further use. After a few failed attempts with other languages/tools (CDO and Julia, see Section 5.1.3), the Python libraries xarray [HH17] and dask [Roc+15] were chosen as the fitting tools for this step. Xarray is a library for handling n-dimensional, labeled arrays. It supports multiple input and output options (amongst others the required NetCDF format) and is compatible with the most popular scientific Python libraries (e.g. Pandas, NumPy). It comes with a great variety

[1] https://docs.dkrz.de/doc/levante/

of features, making it easy to index and transform data and dimensions, joining different datasets (either along a dimension like time or multiple different variables having the same dimensions), and many more. But, most importantly, it leverages the Dask library, which enables xarray to actually use the infrastructure of the DKRZ HPC cluster. Dask is enabling parallel and out-of-the-core[2] computing for the Scientific Python stack. Its goal is to be a NumPy clone leveraging the full potential of modern hardware, which usually utilizes multiple computing cores, without the need for rewriting the already existing scientific Python stack. It uses an acyclic task graph, which distributes tasks efficiently over multiple workers, which can be either different threads or processes. [Roc+15]

5.1.2 Process

In the following, the process for handling one timescope of a member of one scenario is described. For one member, the different timescope files are handled iteratively. Scaling it up for all members is trivial by running them in parallel on multiple nodes of the cluster (for the relatively computation-heavy IVT calculation) or by running different members iteratively (for simple variables). The high resolution of 6 hourly data is used at first for all the variables since it can be trivially reduced to daily/monthly means later. Due to the limited scope of this thesis, only the extreme scenarios SSP126 ("Sustainability: Taking the Green Road" - representing the optimistic development of climate change) and SSP585 ("Fossil-fueled Development" - the pessimistic variant) are compared.

1. Loading the Dataset

In the first step, the process is to load the required dataset(s) into xarray. This means not actually loading the grids into RAM but rather loading the metadata. The actual loading and computation are only performed when required (e.g. when writing the result), every step in between only returns another xarray (meta)dataset. Xarray offers different methods for either loading one dataset file or multiple, the latter is needed for the IVT calculation since multiple different variables need to be used. Important choices here are setting the `compat` parameter and choosing the chunking for Dask. The first one needs to be set to *override*, which prevents xarray to check variables with the same label for compatibility (in this case here dimensions like *lon, lat, time* and the pressure fields *ps*), which is useful e.g. when using dataset of different sources, but since all datasets conform to the same resolutions it is unnecessary.

[2] This usually means handling datasets larger than RAM, using disks (usually SSDs) as extension for RAM.

Figure 5.1: Example of the general overview of the dask dashboard used for analyzing the efficiency of the process.

The latter choice is far more important: In NetCDF datasets, data is often grouped into chunks of a specific, useful size, and these chunks can then be compressed to reduce disk memory usage. If the chunks are compressed reading anything smaller than a chunk is useless, since the whole chunk needs to be loaded anyway to decompress it. Also, reading too small chunks reduces the efficiency of dask, since the introduced scheduling overhead per task is too large and becomes overwhelming. Furthermore, reading too small chunks results in too much worker-to-worker communication, which also results in significantly decreased execution time. On the other hand, reading too large chunks results in memory spills[3] or workers crashing, which both significantly reduce execution time. To find the sweet spot of dask chunk size, dask offers a handy dashboard that visualizes the process of executing the dashed task graph (see Figure 5.1). The gray areas in the *Bytes stored per worker* section in Figure 5.1 show memory spilled to disk, which is an indicator of too large chunk sizes. The red sections in the *Task Stream* section refer to worker-to-worker communication, which may be an indicator of too small chunk sizes if they dominate the *Task Stream*. So, the task overview given in Figure 5.1 indicates a slightly too large chunk size, as too much is spilled to disk. [Buc]

The chunk size in the available datasets is $(192, 96, 47, 1)$[4], which means one chunk corresponds to one time snapshot of the whole atmosphere. Following the previous argu-

[3]This refers to a function in dask where overloaded workers save data on disk to prevent the worker from crashing

[4]Referring to $(lon, lat, lev, time)$

mentation, the only useful way to change the chunks is to use different amounts of time snapshots per chunk. Using the dask dashboard to evaluate different chunk sizes, the optimal chunk size with minimal spilled memory, worker communication, and no crashing workers[5] was $(192, 96, 47, 128)$[4].

2. Reduction to Area of Interest

The next step is to cut out the geographical area of interest, which is the northern Atlantic and Europe. Following [Vie+21a] and [Hur+03], it was defined as $90^\circ W - 40^\circ E$, $20^\circ - 80^\circ N$. Unfortunately for this case, the longitude coordinate is saved in the range of $[0, 360]^\circ$, so it cannot be loaded as one slice. Therefore, the longitude coordinates are first transformed into the form $[-180, 180]^\circ$, with negative values being $^\circ W$ and positive values being $^\circ E$. Then the area of interest can be cut out without problems and the result can be used for further calculations (Step 3) or directly saved as a NetCDF file (Step 4).

The size of the data could be further reduced in this timestep by selecting only the relevant winter months (see Section 5.2), but with future work in mind the whole year was kept in this stage.

3. Calculating the IVT Field

The first step to calculate the IVT field is to convert the hybrid sigma pressure levels (see Section 3.3.2) into actual pressure values. For this, Equation 3.3 is used to calculate a new variable `plev` containing the pressure values at each point in the grid in each time step. Then NumPy's trapezoidal integration function is used to calculate the zonal (Equation 4.3) and meridional (Equation 4.4) components of the IVT. Similarly to the related literature [Aya+22], a constant value for gravitational acceleration $g = 9.806\ ms^{-2}$ is used in the calculation. Using the result of the zonal and meridional components, the norm field $\|IVT\|$ can be calculated using Equation 4.5.

4. Saving Results as a NetCDF Dataset

The results of these calculations (or the geographical box cutout) are then again saved as NetCDF files, in the far less complex directory structure `time_resolution`, `variable`, `member` and then the actual file `timescope.nc`. In case of the IVT, both (zonal and meridional) components are saved alongside the Euclidean norm.

[5] For one of the DKRZ HPC cluster's nodes with 100 GB RAM

5. Generating daily/monthly means

Since the related literature does not entirely agree on the timely resolution of IVT in EOFs (see Table 4.1), the six hourly datasets are reduced to monthly and daily means using CDO. Monthly IVT is also called stationary moisture transport and dominates total water vapor transport [ZY05], while anomalies (departures from monthly mean per daily / sub-daily timestep) are transient components. Both are important parts of total moisture transport, but since a comparison of stationary and transient components is beyond the scope of this thesis, only monthly data will be used for the analysis. The higher resolutions are still saved and can be used in future work.

5.1.3 Challenges of Preprocessing

The steps described in the previous section were just the final attempt. The first idea was to use Climate Data Operators [Sch24], a command-line tool containing multiple operators to process climate and similar data. The operators consist of common statistical and mathematical functions (mean, add, sum), sampling and data selection tools (select geographical or time limits) and other helpful operators like interpolations and even EOF calculation. Although this sounded very promising, it quickly turned out to be very complicated to implement the desired vertical integration in CDO. The following idea was to implement the IVT calculation in Julia [Gao+20], using only a NetCDF library [Bar24], while the rest was coded from scratch. The algorithm was very simple:

1. Load all datasets into the RAM (as recommended by the NetCDF library itself) and cut out the used geographical limits. This should be feasible since all in all one dataset for one timescope-file accounts for $\sim$ 12 GB[6], so the maximum is around 36 GB, since the surface pressure data is not that large ($\sim$ 260 MB)
2. Calculate the IVT with trapezoidial integration multithreaded by handling one timestep by one thread
3. Write the results (Euclidean norm and the meridional/zonal component)

Although Julia promises high performance, it performed quite poorly on the HPC. The reason for this is the slow IO on the cluster: While the calculation itself took only $\sim$ 235 s ($\approx$ 4 min)[7], the loading of the required datasets took around $\sim$ 3350 s ($\approx$ 55 min). This results in roughly 5h (including saving the data to disk) for one member of ScenarioMIP, which leads to 250 h node hours for one scenario. Taking into account that it needs to run

[6] 70 $lon * 32\ lat * 47\ levels * 29220\ timesteps * 4\ byte \approx 12GB$

[7] Referring here and in the following to one timescope of 20 years in one member

for historical simulations as well as other scenarios, this was not feasible according to the limited node hours provided[8].

To reduce the loading time of the data, multiple optimizations were evaluated. First, the amount of data moved in memory was minimized by pre-allocating the needed RAM and writing directly to the pre-allocated space. Furthermore, other NetCDF libraries were tested, but the simple loading times were very similar. Although this significantly reduced the number of allocations, the effect on loading time was negligible. To actually archive a significant increase in loading time, the parallel loading of different files (for humidity, zonal, and meridional wind components, respectively) was evaluated. Unfortunately, the library used [Bar24] encountered a segmentation fault using multiple threads, so the alternative libraries NetCDF.jl and HDF5.jl were explored, since the HDF5 standard allows parallel access to files [Fol+11]. Although parallel access to files using multiple threads (with HDF5.jl) leads to higher speeds in the tests, the results did not yield any significant increased efficiency on the cluster itself. Even splitting up the loading according to the chunking in the files (all data from one timestep is one chunk) and loading each timestep separately in one thread even increased the data loading time quite far. The next approach was to split the task into different processes, each one loading data from one variable. This actually reduced the time spent on tests to one third, but testing it on actual data sizes revealed that the 12 *GB* are too much to be returned from the child processes loading the file to the mother process.

From here on some other approaches could have been explored, like splitting up different time steps amongst different processes, but the far more suitable method of using xarray and dask has been found and implemented.

5.2 EOF Calculation

In the next step, the patterns are calculated using Empirical Orthogonal Functions. While Section 2.2 gives the theoretical mathematical background, this Section describes the practical implementation, the sliding window approach, and domain-specific challenges of calculating EOFs for different variables. The following description is for one member of the different scenarios, but is handled the same way for every member.

This step was implemented in the Julia programming language [Bez+17]. Although it was tried using it for preprocessing and it failed (see Section 5.1.3), the reduction of the data was enough to make it easy to work with the NetCDF library [Bar24] implemented in Julia. Julia has the advantage of making it particularly easy and intuitive working with

[8] Also taking into account that the processes may need to run multiple times due to errors

matrices and high-dimensional arrays. Additionally, it is (usually) also very computationally efficient to do so, which is the reason why it was chosen for the implementation of this step. Other factors are the great reproducibility of the code and the fact that the Makie framework (based on Julia) was chosen for the visual analysis step (see Section 5.3), which made it possible to reuse some already written code.

In general, the procedure for this step is based on the approach of [Vie+21a], but has differences in a few places.

1. Preparation of the Datasets for Sliding Window Approach

In the very first step, the monthly data in the form of multiple NetCDF files containing different time scopes are loaded. Since the influence of the NAO especially significant in the boreal winter, this thesis focuses on the extended winter season like [Vie+21a], keeping only the months December, January, February and March (in the following DJFM). So during loading the irrelevant months are filtered out, in addition to filtering out double values of months (a result of the monthly mean process with CDO), resulting in a timeline containing exactly one spatial map for each month value for each month. The data is now available in the form of a three-dimensional array of shape $(lon, lat, time)$, while longitude, latitude, and time are stored separately as one-dimensional arrays. To make changes from the (pre)industrial time to future scenarios visible, each scenario is concatenated with the historical simulation. The data is then collected in certain scopes for the sliding-window approach. First, the data is grouped into winter seasons, one season is defined as a slice of data before a huge time threshold (e.g. 150 days, representing the spring/summer/autumn gap) to the next date. Then, a scope is defined as a specific number of such winter seasons. Following the argumentation of Vietinghoff et al. [Vie+21a], the window size affects the smoothing of the data: The larger the time window, the better the noise is smoothed out. However, as a drawback, small-scale features are smoothed out along the noise. But since the focus of this thesis is monthly mean data, we are interested in the large scale structure of IVT and its connection to other variables, so rather large windows of 30 and 50 years were chosen, similar to Vietinghoff et al. [Vie+21a]. The next scope is then shifted by one winter season. In contrast to [Vie+21a], the winters themselves are not reduced to one mean map, since this has not been done by any related work using IVT fields, and smoothing out the data any further could potentially remove important features.

From here on, steps are described for one time window of any variable. The same steps are then applied to any other time interval for any instance of any variable.

2. Preparation of the Data for SVD

Before calculating SVD, the data must be prepared accordingly. After scoping, the data for each scope are available as a three-dimensional array with the shape *(lon, lat, time)*. Following the argumentation from [Vie24] and Section 2.2 the first step is to multiply each datapoint with the factor $\sqrt{m-1}^{-1}$[9]. Then the geographical weights are applied, shifting the data into weighted space. The geographical weights are applied depending on the latitude of the data, approximated by $\cos(lat)$. Next, the data needs to be reshaped, since SVD only works for matrices (two-dimensional arrays): The geographical dimensions (lon, lat) are reduced to one spatial dimension and then permuted, so the time is the first dimension. Now the shape of the data is $(time, spat)$, and SVD can be calculated.

3. Calculating EOFs with SVD

In the next step, the singular value decomposition is applied to the geographically weighted two-dimensional data chunk, using the SVD implementation of Julia's `LinearAlgebra` package (which is part of the Standard Library). This approach computes all m[9] modes, although only the top five modes are saved. This was feasible for the monthly data used in this thesis, but once this approach is scaled up to daily or sub-daily data it should be replaced by faster algorithms like Snapshot POD or the SLEPC implementation used by Vietinghoff [Vie24]. Those implementations compute only the first n modes, which is significantly faster than the full SVD computation.

The result of the SVD of the weighted data chunk D is then:

$$L, S, R = SVD(D) \tag{5.1}$$

L are the left singular vectors, R the right singular vectors, and S a vector containing the singular values. With time being the first dimension as a result of the previous step, R contains the spatial modes, L the temporal modes (also called principal components or EOF coefficients). The singular values Σ can be used to calculate the eigenvalues (σ_i^2), which are the share of variability encoded in that corresponding mode i with Equation 2.11. Then the modes are cut of at a limit n (usually five) to reduce the usage of disk space and the loading times in the visualization processes later.

[9] m being the time dimension size

4. Post Processing of the EOFs

After the first n modes of interest were cut off, some more steps are required to reconstruct the anomaly map generated in the preparation step. First, the temporal modes are scaled with $\sqrt{m-1}$, then the modes are aligned with some vector (see Section 5.2.2) and the weighting is reversed by multiplying the spatial modes depending on their latitude with $\cos(lat)^{-1}$. The results are then saved using a data structure serialization library called JLD2, which is based on HDF5 [24]. The five most significant spatial and temporal patterns are persisted, along with the singular values and the sum of all eigenvalues (for scaling and variability computation). As stated in Section 2.2, the spatial as well as the temporal patterns from the SVD are in unit scale, and since they are eigenvectors, any combination of a scalar c and EOF modes $cg^{(k)}$ and their corresponding temporal pattern $\alpha_k^{(j)}c^{-1}$ are also viable solutions to the problem. This leaves room for the scaling and choosing the sign of the spatial/temporal patterns in a way that simplifies interpretation of the patterns. The former is explained in Section 5.2.1, and the latter in Section 5.2.2.

5.2.1 EOF Mode Scaling

As explained by Vietinghoff [Vie24], there are two particularly useful ways of scaling the results: Either by multiplying the temporal patterns (here: the left singular vectors) with their corresponding singular value σ_i, which yields EOF coefficients in the original unit of measurement, or by multiplying the spatial pattern (or right singular vectors) with their corresponding singular value σ_i, which yields maps of varibility patterns in the original unit of measurement. Depending on the type of visualization used, the former or the latter is more fitting, so in contrast to [Vie24], scaling is done during data loading to be more flexible.

5.2.2 EOF Mode Alignment

Since basically only two scaling modes are useful (see Section 5.2.1), the sign/direction of that vector is the last but very important choice to make. Although there is no inherent meaning in the sign of one certain pattern, it is crucial for understandability to align the results to a) compare modes across time and members and b) analyzing spatial and temporal patterns separated from each other. Most of the related work in Section 4.3 uses a maximum of a few modes (since most don't compute multiple modes, and none use an ensemble scale simulation), which enables them to align the modes by hand to be useful or don't align them at all. Since this work has many more patterns to align for one scenario, it is not feasible to do it by hand, the problem needs to be solved algorithmically. The only

work using such a large-scale analysis is by Vietinghoff et al. [Vie+21a]. Its solution for this is providing a field F to which we compare our spatial mode and decide whether to flip it ($\equiv$ multiplying it with -1) or not.

The method chosen by Vietinghoff [Vie24] is to use the scalar product of the spatial pattern with the mean field used to generate the anomaly maps in the data preparation step. This mean field is then adjusted by the spatial mean, to reduce it to the actual values of variance. If the result of the scalar product is less than zero, which means the spatial pattern is closer to the mean of the data. This has the effect that positive values of patterns align with above-average values of the actual data mean and vice versa, making the sign of the EOFs interpretable [Vie24].

Unfortunately, this works only well for the primary EOF mode, while the rest compared to the adjusted temporal mean yields not so successful results. A way of testing the alignment of patterns is by computing the cross-correlation boxplots from Section 5.3.2 with and without absolute correlation values. If the absolute correlation boxplot shows a clear trend (e.g. consistent values) but the plot with normal correlation values has either no visible correlation (meaning the alignment is very arbitrary and varies a lot) or there are many outliers on the opposite side of the zero line (alignment did not work for just some members). To fix these problems and also measure correlation of the secondary or lesser modes, the spatial pattern of the first scope in the historical simulation is used as the field to align to as the best effort approach. Depending on the mode and variable, this was more or less successful. The top five spatial patterns of each relevant variable (IVT, surface pressure and precipitation) are shown in Figure 5.3.

5.3 Analysis of EOF Patterns

After generating and storing the EOF results, they can be used for analysis. This section contains the techniques and reasoning for decisions, while Chapter 6 contains the descriptions and analysis of the full results. This section also refers to techniques used in related work.

For this task Julia's Makie framework [DK21] was chosen. It allows one to easily create a complex layout for figures and allows one to create animations as well as interactive visualizations. It was made to "create high-performance, GPU-powered, interactive visualizations, as well as publication-quality vector graphics with one unified interface" [DK21]. In addition, a library[10] is available to project data on different map projections. Unfortunately, this library turned out to be not fully mature, leaving the Mercator projection (see

[10] GeoMakie.jl: https://geo.makie.org/

Figure 2.2) as the only working projection, since others do not support cutting out specific geographic regions.

While the previous two Sections had the goal of generating spatio-temporal EOF patterns (**M1** from Section 1.3), this Section has the goal of fulfilling the other two milestones: Studying the relationships with other variables (**M2**) to get a sense of the meaning of IVT EOF patterns and display the variability introduced by the multiple members of the chosen dataset (**M3**).

5.3.1 Spatial Pattern Analysis

The main focus of this thesis lies on the visualization of the spatial patterns, as those are easy to understand visually. The main challenge here is to visualize the evolution of the pattern over time while still encoding the variability originating from the 50 members of the MPI GE CMIP6. Of course, just one member could be used, or the data can be reduced to averages, but this also means losing the advantages of multiple member simulations. Although there are some examples of visualizing uncertainty in scalar fields (or features thereof like contour lines), none of them provide an existing implementation, which greatly hinders the usability. Since a new implementation of these concepts was outside the scope of this thesis, a new way of visualizing the ambiguity of multiple members needed to be found. Although there are ways to represent uncertain scalar fields (such as [Con+11]), it may not always be helpful to visualize the full ambiguity of the data set, as it could generate too much visual clutter. Instead, a certain feature could be extracted to show the variability of the ensembles.

Feature Selection

As shown in Section 4.4, there exist multiple ways of visualizing scalar fields, either by transforming the actual scalar field in some way (like the approach of Coninx et al. [Con+11]) or by selecting an interesting feature in it and visualizing it. The latter approach was used by Vietinghoff [Vie24] (using the extreme points of the fields) and by [San+10; WMK13] using contour lines.

After analyzing the different spatial patterns of the different variables (and especially IVT), it was obvious that the separation line between the positive and negative spatial patterns seems to be very pronounced, especially in the primary modes. Therefore, the use of contour lines seemed like a good choice for a feature. Contour lines (also referred to as isolines) are lines that share the same value of the function that defines the field. The best known contour lines are lines of the same height in mountains maps. Since

one question of the introduction was how those spatial patterns shift geographically, the contour line of zero, representing the borders of positive and negative patterns, seems fitting. This procedure relates to the idea of level crossing probabilities, similar to the work of Poethkow et al. [PPH13], but for contour lines and not for isosurfaces.

Visualizing the Feature

The most straightforward way of visualizing uncertain contour lines is spaghetti plots as used in the work of Sanyal et al. [San+10], simply drawing all the isolines of the different members, usually in different colors. This was also implemented in this thesis, but spaghetti plots have quite a few drawbacks. While they work quite well when the results of all members align properly and do not differ that much, they quickly get chaotic once that is not the case. This also scales with the number of members in the ensemble. A spaghetti plot with 10 members may look fine but looks quite chaotic with 30. Once members are not easily countable anymore, it is hard to instinctively differentiate between a quite uncertain area (10 out of 50 members) and a quite likely area (25 out of 50 members). To reduce this and make it possible to instantaneously evaluate how likely an area is to contain isolines, the hexbin approach was explored. Additionally, contour lines give a false sense of precision in data that is actually not that precise.

Level Crossing Probabilities using Hexbins

To fix the spaghetti plot issues, another method using hexbins was implemented. Hexbin plots[11] as used by Carr et al. [COW92] for geographic data are an alternative to heatmaps, showing the density of observations (usually given as points in space). Heatmaps divide the observed grid into rectangular areas of variable size, and colors the rectangular area depending on the number of observations in it. Hexbin plots are very similar, except that they divide the grid into hexagonal "bins" and handle the observations in the same way as heatmaps. The advantage of hexbin plots is that it represents the distribution better than square bins (≡ heatmaps) as depicted in [COW92], and that it can be more visually appealing to humans [COW92].

The threshold for displayed hexbins was set to one, so that areas without any contour line stay free of hexagons. The preparation of points was conducted as follows: First, the required contour lines were computed, which are represented as a list of 2D points. Unfortunately, the resolution is too high at some places (multiple observations per bin) and too low at others (no observation hitting the bin) as shown in Figure 5.2. The solution for this

[11] https://docs.makie.org/dev/reference/plots/hexbin

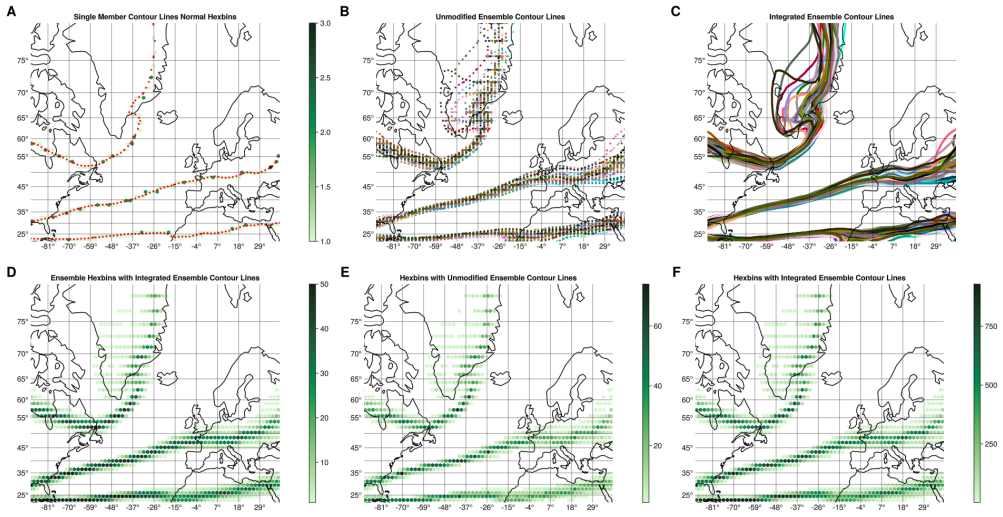


Figure 5.2: Example of the hexbin approach using the dominant IVT pattern. Sample distance is 0.1. A are the hexbins with one member. B and C are the contour line vertices, not integrated and integrated, respectively. E and F are the corresponding normal hexbin plots. D are the presented ensemble hexbins applied to the integrated countour lines in C.

involves two steps: For the first problem, a Makie recipe[12] was implemented that accepts grouped observations, allowing only one observation to count for one hexbin per group. For the second, a sampling algorithm was used to sample along the line. With a sufficient sampling distance, no bin is missed no matter how small the hexbins are. Therefor, the range of possible values is one (hexbins with value zero are not plotted), only one member's contour line passes through that bin, representing an outlier, to 50, meaning that all the contour line is very certain at that point. An analysis comparing this approach with traditional spaghetti plots is given in Section 6.3, while in Section 6.1.2 the interpretation is explained on the actual results.

5.3.2 Temporal Pattern Correlation

The spatial pattern is, of course, only half of the EOFs. An equally important part are the EOF coefficients, representing the activity of the spatial patterns in each (monthly) time step. Unfortunately, it is very hard to visualize them on ensemble scale, since each member is quite different in each month, resulting in a hard-to-interpret, visually cluttered plot when using mean functions or boxplots.

[12]This is the name for certain types of plots in Makie, like surface, heatmap or boxplot.

So, instead, the temporal activity can either a) be analyzed on its own using standard deviation or b) it can be used to evaluate the relationship of certain patterns with other variables or even other patterns. For the former, it is important to scale the EOF coefficients with the singular values since on their own they have a standard deviation of one. The standard deviation of the scaled EOFs is then calculated for each sliding window and member and visualized using boxplots.

For the latter, it is very important that the same members are compared across variables since they share their initial conditions and forcings. The tool of choice for comparing the temporal patterns/signals is (cross-)correlation, which measures the linear relation between two signals. This does not imply any causality in any direction on itself, but shows how similar signals (e.g. the EOF coefficients) are. Causality must therefore be justified separately from correlation. Cross-correlation measures the correlation between two signals with a predefined range of lags (e.g. $-n, \dots, 0, \dots, n$) to find correlations which are shifted in time. A lag of zero is then equal to the usual, non-shifted correlation.

Comparing two EOF patterns

To explore the relationships between certain patterns, the EOF coefficients can be compared with the EOF coefficients of another variable. This can answer the question of which patterns share activity in a certain month and how this relationship evolves over time.

This is evaluated using cross-correlation. The reason for this was that by introducing a lag in the signal, it could reveal certain temporal relationships, e.g., a positive IVT EOF coefficient of EOF2 (see Figure 5.3) leads to positive activity of the precipitation pattern in Great Britain (e.g. PR EOF2).

This evaluation was carried out as follows: per scope (30 or 50 winters), the cross-correlation of two different variables X and Y was calculated for two modes a and b by calculating the cross-correlation of their temporal patterns $c_a^X(t)$ and $c_b^Y(t)$. The range of lags used was $[-n, n] \in \mathbb{Z}$, n being half of the scope length. Of course, a greater extent of lag could be chosen, but it seemed very unlikely that moisture transport in a certain month would affect precipitation many decades later (the same reasoning for the other variables). Then, the maximal extent of the correlation per member was used for a boxplot of that scope. The lag associated with that value per member is then displayed in a separate boxplot of lags. Consequently, this procedure was repeated for every scope, from the beginning of the historical simulation to the end of each scenario, depicting the evolution over time.

Since only temporal coefficients (without spatial patterns) are evaluated here, it is crucial that the alignment of the patterns (as described in Section 5.2.2) works well, as it can

significantly distort the correlation results. In fact, using this kind of analysis revealed the lack of stability of the procedure used by Vietinghoff et al. [Vie+21a] in EOF2 and lesser modes. For this, the plots of the maximal absolute value of correlation were compared with the normal maximal extent of correlation. If the plot of absolute correlation revealed a consistently high value, but the normal correlation fluctuated around zero (or had many outliers mirrored on the x-axis), the alignment of patterns did not work correctly.

Comparing an EOF pattern with another variable

Since it is notoriously hard to connect (mathematical) EOF modes with real physical modes [DL02; HJS07], it is not enough to analyze the relationships between patterns since they may not represent any actual physical modes (see the simple example explained in Dommenget and Latif [DL02]). Instead, the recommendations of Dommenget and Latif [DL02] were followed, using different statistical tools to evaluate the modes. Some of the related work [LZ12; Zou+18; ZY05] used regression maps, depicting the regression slopes between the EOF coefficient $c_a^X(t)$ of the variable X and the mode a (e.g., the dominant EOF mode of IVT $c_1^{IVT}(t)$)and the temporal evolution of each gridpoint $Y(lon, lat)(t)$ of another variable Y. Fernández et al. [FSZ03] used a similar procedure but with the correlation of $c_a^X(t)$ and $X(lon, lat)(t)$, illustrating the correlation coefficient for each gridpoint with the temporal pattern for the same variable X. This addresses the problem described by Dommenget and Latif: "The PCs of the dominant patterns are often a superposition of many different modes that are uncorrelated in time and that are often modes of remote regions that have no influence on the region in which the pattern of this PC has its center of action." [DL02]

For this thesis, an interactive comparison of the patterns of any variable X and the actual data of another (or the same) variable Y was implemented. To display the ambiguity introduced by the different members, spaghetti plots and the hexbin approach described above were reused, highlighting contour lines of certain correlation levels. Therefore, they display areas with a higher (or lower) than an interactively chosen correlation level (e.g., 0.7). Also, the mode and scope can be interactively chosen, to explore different modes and their evolution in time. Of particular interest here is the correlation between IVT modes and precipitation data, which could help to explain how the dominant IVT modes influence precipitation in Europe.

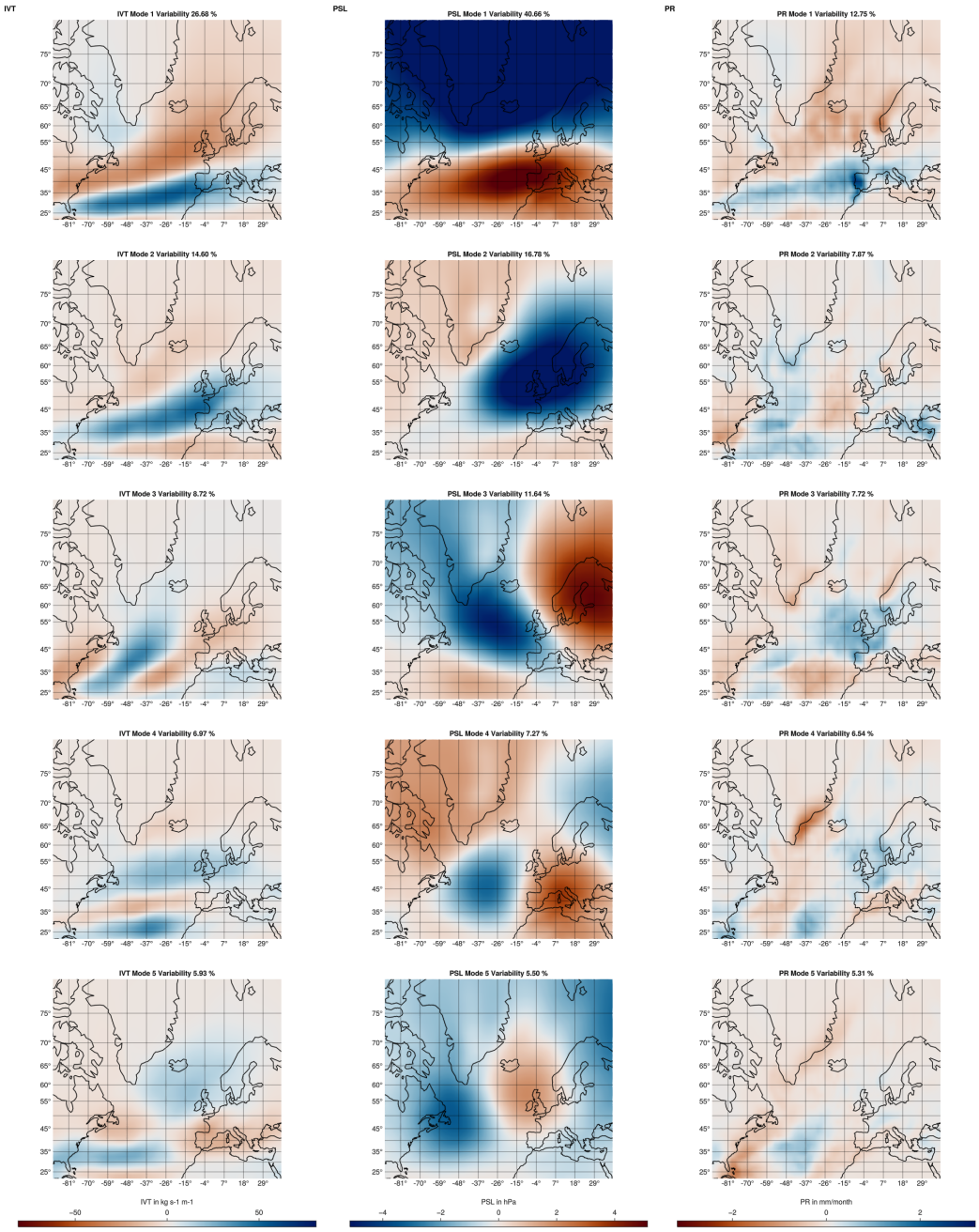


Figure 5.3: The five most significant modes of each of the processed variables (using the short names given in Table 3.1). Taken from one member with 50 winter timescope and the scope from 1868 to 1918

6 Results

This chapter presents the results of the analysis described in the previous chapter. Comparing the results of 30 and 50 seasons per timescope revealed (similar to [Vie+21a]) that using 30 winters is, in general, more noisy, but the results are structurally very similar. Therefore, only the smoother results of 50 winters per timescope are shown in this section.

6.1 Evolution of Patterns

This Section gives an overview how the EOF patterns change over the time, also comparing the differences of the two chosen climate scenarios, which represent the extremes of climate change handling.

6.1.1 Evolution of Encoded Variability

The first simple evaluation is to look at the change in the share of variability encoded by each EOF (see Equation 2.11). The results are displayed in boxplots, with the colored bar being 50% of the members. The whiskers are 1.5 the size of the interquartile range (distance between upper and lower and of the colored bar), any data point outside that is considered an outlier and represented with dots.

Figure 6.1 shows that there is no striking change in the SSP126 scenario in any way. The five most significant modes stay pretty much the same across the studied 250-year time period, with the primary mode (NAO) encoding around 39% (median) of the whole dataset variability in each time scope, with fluctuations of the interquartile range (50% of the data) introduced by the simulation members being around ±2%, with no significant trend over the years. The median of the secondary mode (EAP) stays around 17%, with the quartiles being ±1%. The median variability encoded by EOFs 3,4 and 5 is around 13%, 8%, and 5%, respectively. Compared to the SSP585 scenario, it is obvious that there is very little or no change in Modes 3-5 and 1. But interestingly, the median variability encoded by the secondary mode rises from 17% in the 1850 - 1900 scope to around 20% in the last, exposing a clear trend over the course of climate change.

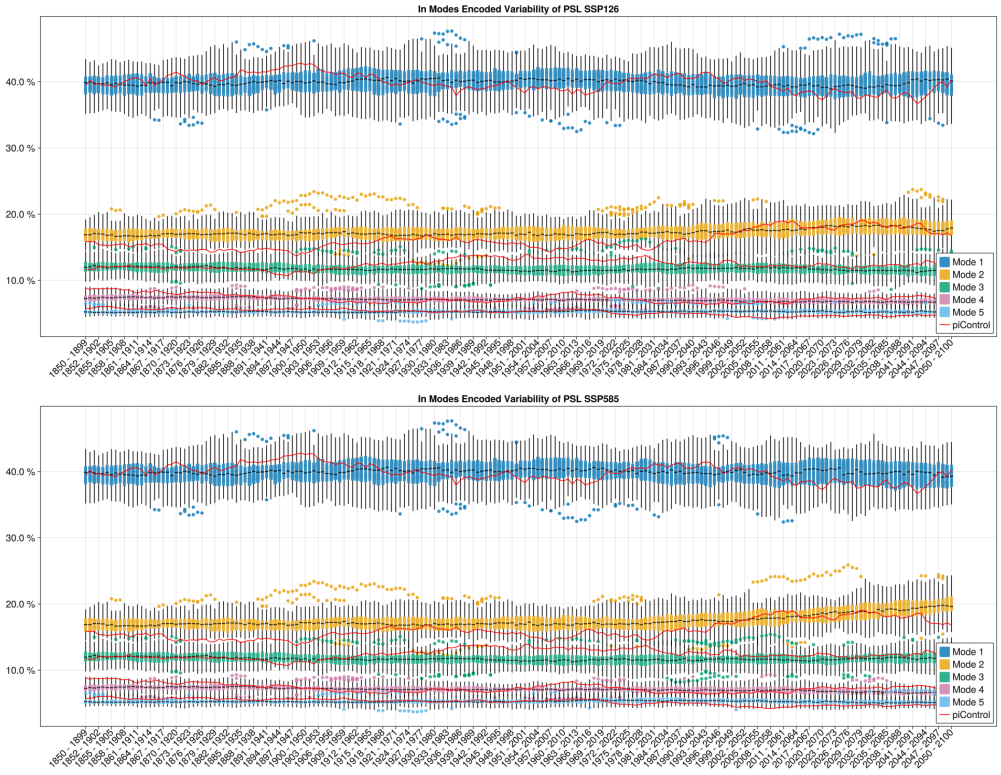


Figure 6.1: Boxplot of the variability encoded in the top five modes of PSL EOF. Red line shows piControl run modes.

The same analysis with the IVT patterns (Figure 6.2) reveal a general upward trend in the primary mode of IVT, from median 26% in the first window to around 28% in the last. This trend is very similar in both SSP126 and SSP 585. Modes 3,4 and 5 also look very similar in both evaluated scenarios, with a median encoded variability of 8%, 6%, and 5%. Similarly to Figure 6.1, the secondary mode (representing around 15% of variability) shows an upward trend in the scenario SSP585 to around 17%, which is not recognizable in the scenario SSP126.

The dominant modes of precipitation EOF seem to account for far less of the total variance (28.5% of the mean total variance of the top three modes at the beginning of the historical simulation), compared to the other patterns (50% and 68% for IVT and PSL, respectively). Comparison of the evolution of the variability of the mode of precipitation EOFs (Figure 6.3) shows no significant changes in modes 3,4 and 5 between the two scenarios evaluated. Those encode on median 5%, 6%, and 6.5% with small fluctuations introduced by the members. Mode 2 also looks very similar in both scenarios, with a median

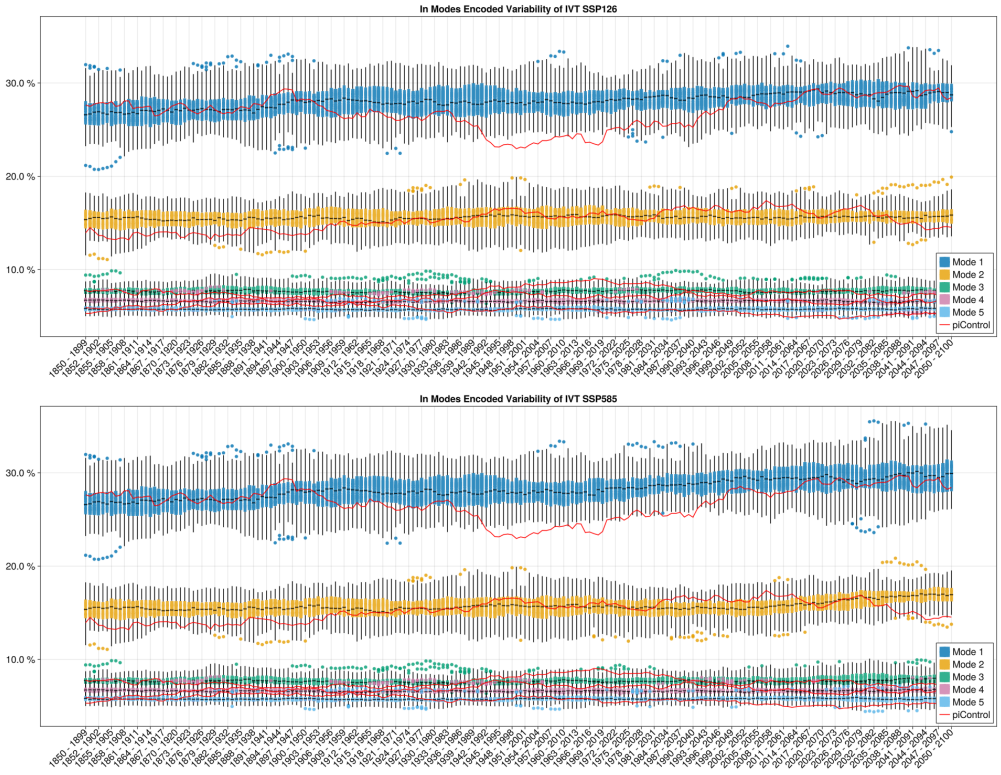


Figure 6.2: Boxplot of the variability encoded in the top five modes of IVT EOF. Red line shows piControl run modes.

encoded variability of around 8.5%. The primary EOF on the other shows significant differences between scenarios: While it has a far greater variability across members than the other modes and follows a general upward trend in both SSP126 and SSP585, it is more pronounced in the latter. It evolves from around 12.5% in the 1850-1900 window to around 14% in SSP126 and 15.5% in SSP585.

In general, the modes beyond the second seem to be not well separated from the others, which is why they will not be analyzed in detail in the following sections.

6.1.2 Evolution of Spatial Patterns

This Section shows the evolution of the spatial EOF patterns, shown in Figure 5.3. Since the EOF modes three, four, and five are generally quite low and similar in their eigenvalues (which directly correspond to the variance (see Equation 2.11) encoded shown in Figures 6.2, 6.1, and 6.3), they are left out of the analysis of this and the following sections, as modes' eigenvalues need to be well separated from each other [HJS07]. As already men-

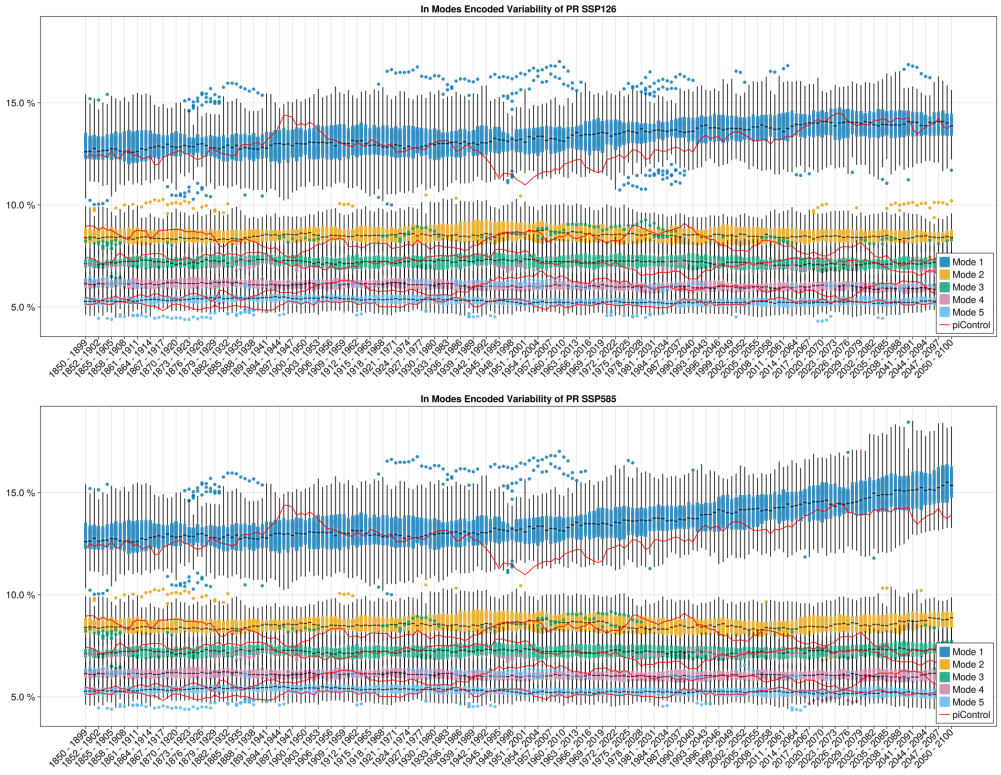
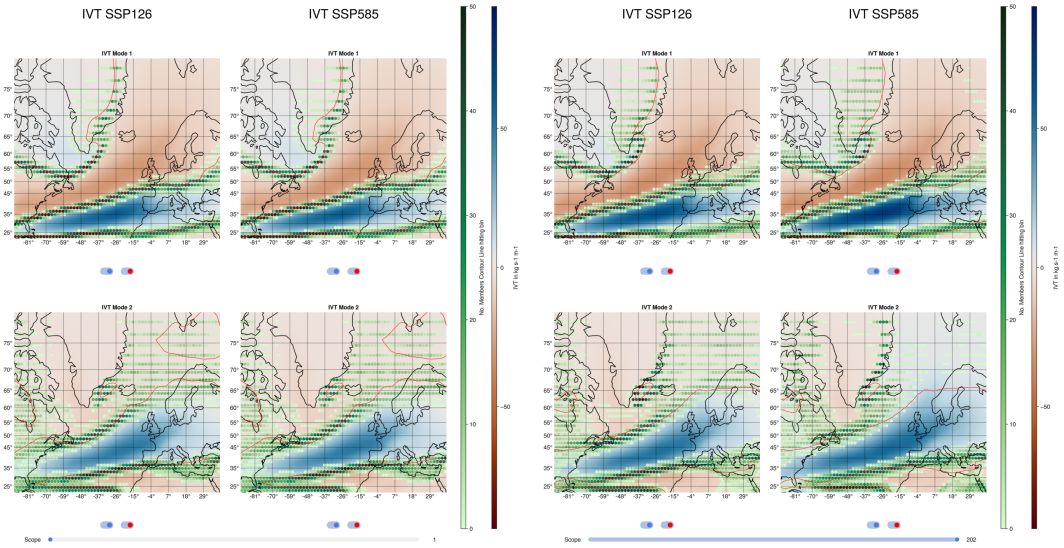


Figure 6.3: Boxplot of the variability encoded in the top five modes of precipitation EOF. Red line shows piControl run modes.

tioned, mode three also seems to be not well separated, but is still shown in the following figures for an example how an inconsistent mode looks (at least for IVT and precipitation), which makes it hard (or even pointless) to evaluate it for an ensemble simulation. Usually, the rule of thumb introduced by North et al. [Nor+82] is used, but since the eigenvalues of the first three modes (or two for precipitation, see Figure 6.3), this is left out of this thesis.

The variability introduced through the 50 members of MPI GE CMIP6 is displayed here with the hexbin approach explained in Section 5.3, while a discussion of the visualization of hexbin compared to classical spaghetti plots is given in Section 6.3. The images in this section are the spatial patterns of the first and last timescopes, which show the evolution of the different scenarios.

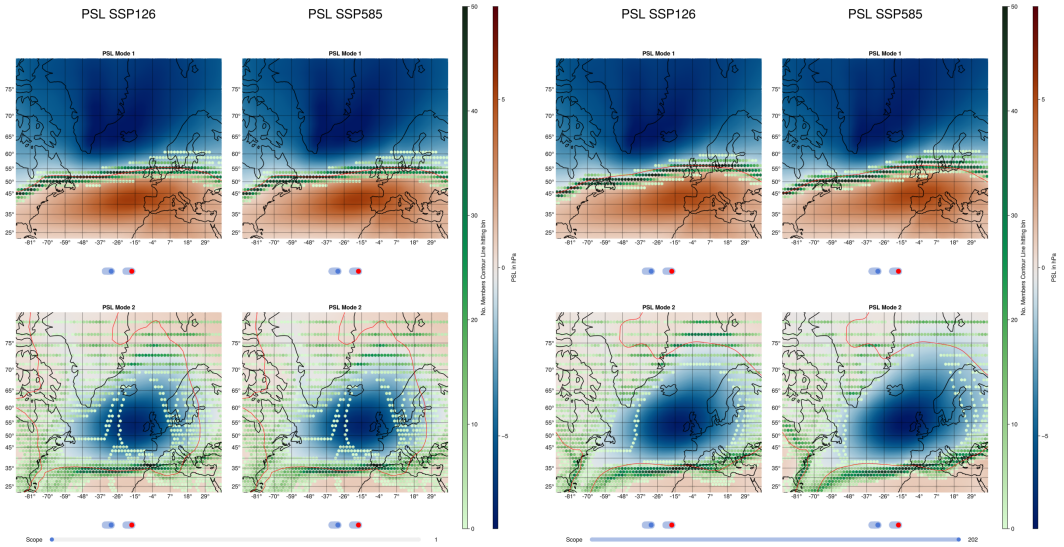
Figure 6.4 shows the evolution of IVT EOF spatial patterns. In general, regardless of the future scenario, EOF1 and EOF2 stays structurally very stable across all the ensembles' members, which can be seen on the clear, dark green borders of the colored surfaces. EOF3 on the other hand seems pretty unstable, since most of the map is covered in light



(a) First scope of historical simulation

(b) Last Scope of scenario simulation

Figure 6.4: The two primary spatial modes of IVT EOF, with a 50 winter scope and hexbins visualizing the variability introduced by simulation members. Figure 6.4a shows the state in the historical simulation (second half of 19th century), while Figure 6.4b displays the state in the second half of the 21st century. The left column in each picture is the result of SSP126, the right of SSP585. The red line shows the contour line of zero of the preindustrial control simulation.



(a) First scope of historical simulation

(b) Last Scope of scenario simulation

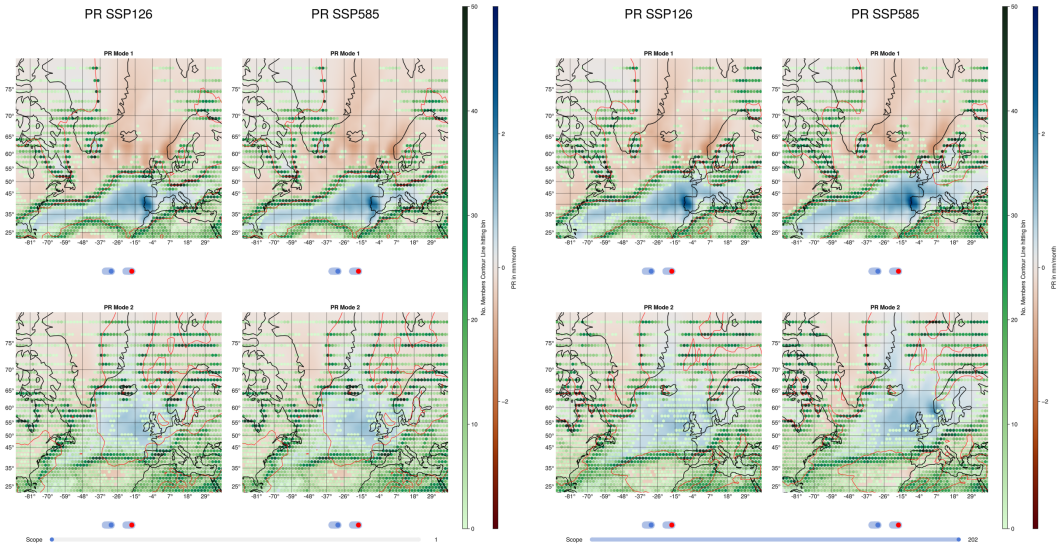
Figure 6.5: Same as Figure 6.4, but with PSL patterns.

greed hexagons, which means that the contour lines of zero switch significantly between all the 50 members of the ensemble. This also has consequences for the alignment across members and time, which will obviously not work if the patterns differ greatly across time and members. Therefor, the analysis regarding such patterns will be kept short since the multi-member, sliding window analysis of such patterns used in this Thesis does not apply very well to such patterns.

The dominant EOF1 pattern of IVT is characterized by a positive IVT values reaching from Florida to Spain and negative values from the east coast of the USA to Northern Europe and the northern Atlantic. There are three clearly visible borders of these positive and negative areas, associated with three groups of contour lines: The first going through Canada, quite coherently across members (many dark green hexagons in a row), and then following the east coast of Greenland, fading out over the mainland quite differently across the members (many light green hexagons in a larger area). The second border follows a similar pattern: Starting quite coherent across members at the beginning of the Florida peninsula, over the Atlantic to the east coast of France, and then fading out differently in eastern Europe. The third border goes through the most southern part of the evaluated area, through North Africa and then to the Arabian Peninsula, staying pretty consistent across its path. Comparing the state of the patterns at the end of the different future scenario simulations, the change is quite subtle, but visible in the area of France and the south of the British Islands: While the dark green hexagons in the beginning of the historical simulation are on/below the red line of the preindustrial simulation, the majority of zero contour lines in SSP126 seem to be above the preindustrial control contour line. In SSP585, the dark green hexagons stretch even further north, indicating that the slight northward shift of IVT EOF1 at the end of the 21st century is even more pronounced.

The IVT EOF2 is characterized by a strong IVT anomaly right were the separation contour line of EOF1 is, and lesser, opposite anomalies in the north and south. Comparing the beginning of the historical simulation with SSP126, the changes seem only minimal, whereas the difference of SSP585 and the others are far more pronounced: While the southern, prominent separation contour line seems to be more variable across members (broader hexbin band), the northern isoline appears to move to the east, opening the pattern up to the north. Also, the big area of fading out contour lines in the north of Scandinavia, seems to now be (more or less) uniformly part of the positive anomaly. While mode three of IVT EOF does expose changes between the different scenarios (again historical simulation and SSP126 look very similar, while SSP585 differs), the patterns seem to be pretty unstable compared to the other, indicated by the whole map being covered in light green hexbins.

This implies that the contour lines of zero are all over the map, which means that the structures are quite different across the simulations' members.



(a) First scope of historical simulation

(b) Last Scope of scenario simulation

Figure 6.6: Same as Figure 6.4, but with precipitation patterns.

The evolution of sea level pressure patterns is displayed in Figure 6.5, which shows the North Atlantic Oscillation and the East Atlantic Pattern (see Section 1.2.2). While the first two EOFs are well known, real world phenomena, the third mode was never mentioned in the related literature. Although it seems to be stable across members and time scopes, it will therefore be left out of the analysis. The NAO in mode one, characterized by its north-south dipole of pressure, is very consistent across the 50 members of the simulation, indicated by the narrow band of hexbins showing the switching line between the positive/negative anomalies. Comparing the historical simulation with SSP585, there seems to be a slight northward shift of the hexbin band, best visible at the northern part of the British Islands. This shift seems to be less pronounced in SSP126. The NAO does not exhibit outliers across all sliding windows. In contrast, the EAP appears deformed in some members, as evidenced by a few very lightly colored hexbins in the blue center of the historical part. These deformations are also observable in other scopes throughout the simulations, which are not related to the historical part of the timeline. The southern border of the prominent negative pattern remains indistinguishable between the scenarios and the historical simulation. However, the stable northern border, marked by dark green hexbins, seems to shift further north, particularly in SSP585.

Similarly to the other figures in this Section, Figure 6.6 shows the evolution of dominant precipitation EOF patterns. While mode 1 looks very similar to the primary pattern of IVT EOF, the secondary mode looks quite comparable to the EAP. Quite noticeable in the primary pattern is the strong positive precipitation anomaly at the west coast of the Iberian Peninsula. Also, EOF1 is associated with negative anomalies along the southern east coast of Norway and a small positive pattern on the southern-western coast of Greenland. The SSP126 EOF1 pattern extends considerably into mainland Europe and also extends further north. This effect is noticeable only in Europe and not in the Atlantic, with the anomaly along the coast of the Iberian Peninsula becoming more pronounced (indicated by a more intense blue). In the SSP585 EOF1 pattern, the effects observed in SSP126 are even more pronounced, with the piControl line comparable to the farthest outliers of SSP585. EOF2 of precipitation appears to be quite unstable across the members of the simulations, as indicated by hexbins filling most of the map. The only somewhat stable area is a positive anomaly over Great Britain and slightly further north. A small but consistent negative anomaly is observed along the east coast of Iceland in many members. However, this stability collapses in some members and time steps, as shown by very light green hexbins moving through the center. In future scenarios, the central positive pattern opens up to the north, resembling the evolution of the EAP. This effect is more pronounced in SSP585 than in SSP126.

6.1.3 Evolution of Temporal Patterns

As explained in Section 5.3, there is no proper way to visualize the EOF coefficients of an ensemble in the same way as in the related work. Therefore, the only way left is to analyze the statistics of those signals and compare them in a boxplot to get a sense of variability across members. In Figures 6.7, 6.8, and 6.9 the evolution of the standard deviation (SD) of IVT, sea level pressure, and precipitation is displayed, respectively. Keeping in mind that being scaled to the original unit means multiplying each EOF coefficient series with the associated singular value, which is very closely related to the encoded variance (see Equation 2.11).

Analyzing the evolution of the SD of the IVT EOF coefficients, that compared to the quasi-stationary preindustrial control simulation the SD increases significantly, regardless of the scenario. While mode 1 experiences a drop in SD in the scopes from $\approx$ 1933 to 1999 (start of the scope), the SD of the members keep steadily increasing. This increase is much more pronounced in the SSP585 scenario than in SSP126, resulting in a median SD of $\approx$ 970 $kg\ s^{-1}m^{-1}$ (compared to $\approx$ 820 $kg\ s^{-1}m^{-1}$ in SSP126). This change is much more extreme than the sole change in proportionate encoded variability shown in Figure 6.2.

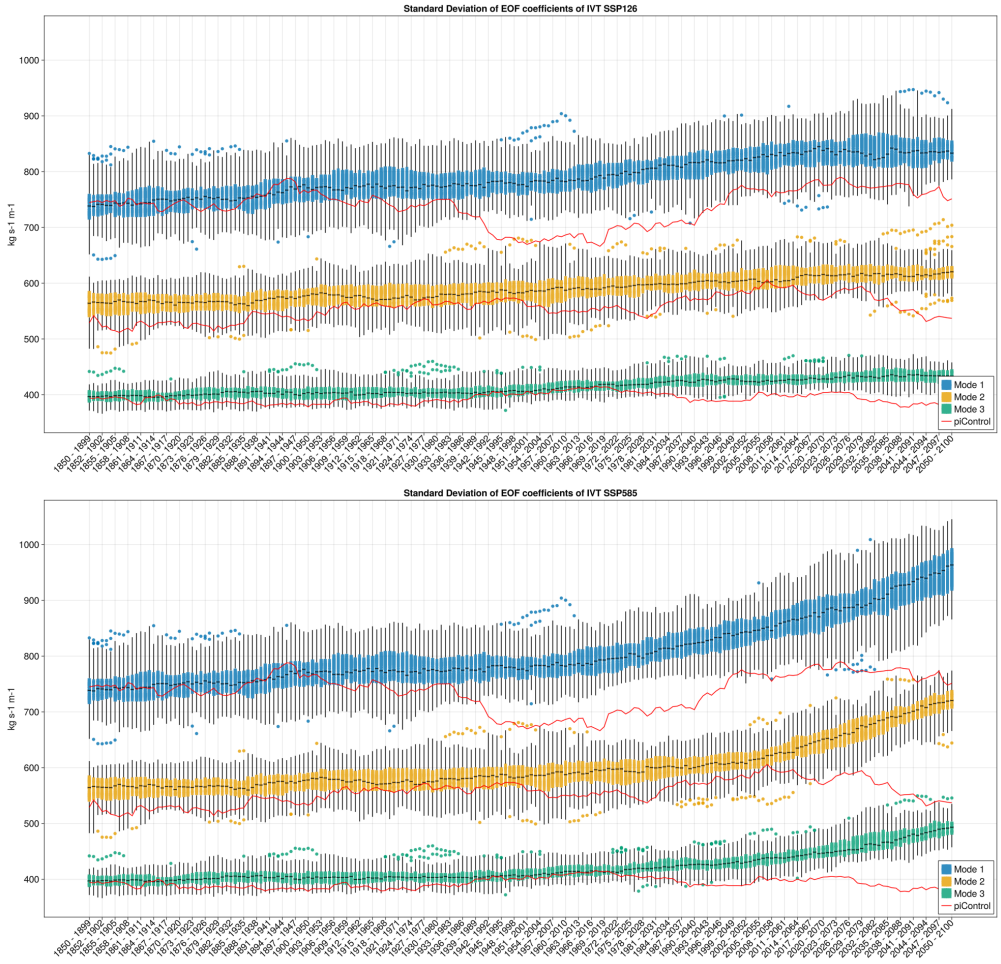


Figure 6.7: Evolution of the standard deviation of IVT EOF coefficients for each scope and simulation member. Top plot shows results for SSP126, bottom plot for SSP585.

The variance across the members seems pretty similar in both scenarios. All of this applies to modes two and three as well: Slow but steady increase in SD, which remains consistent in SSP126 but experiences a sharp increase in the later scopes of SSP585.

Regarding the evolution of SD of sea level pressure EOF, it is obvious that the levels are more consistent across all scopes, regardless of the scenario. Also, the piControl simulation is completely in line with the corresponding boxplots and could not be distinguished from the historical/scenario simulation. When comparing the scenarios, the SD of the Mode 1 coefficients (NAO indices) seem to be very similar to each other, with SSP585 having maybe a bit more variability across members. While it is also hard to distinguish the results of the different scenarios for Mode 3, SD evolution of Mode 2 (EAP indices) seem to experience

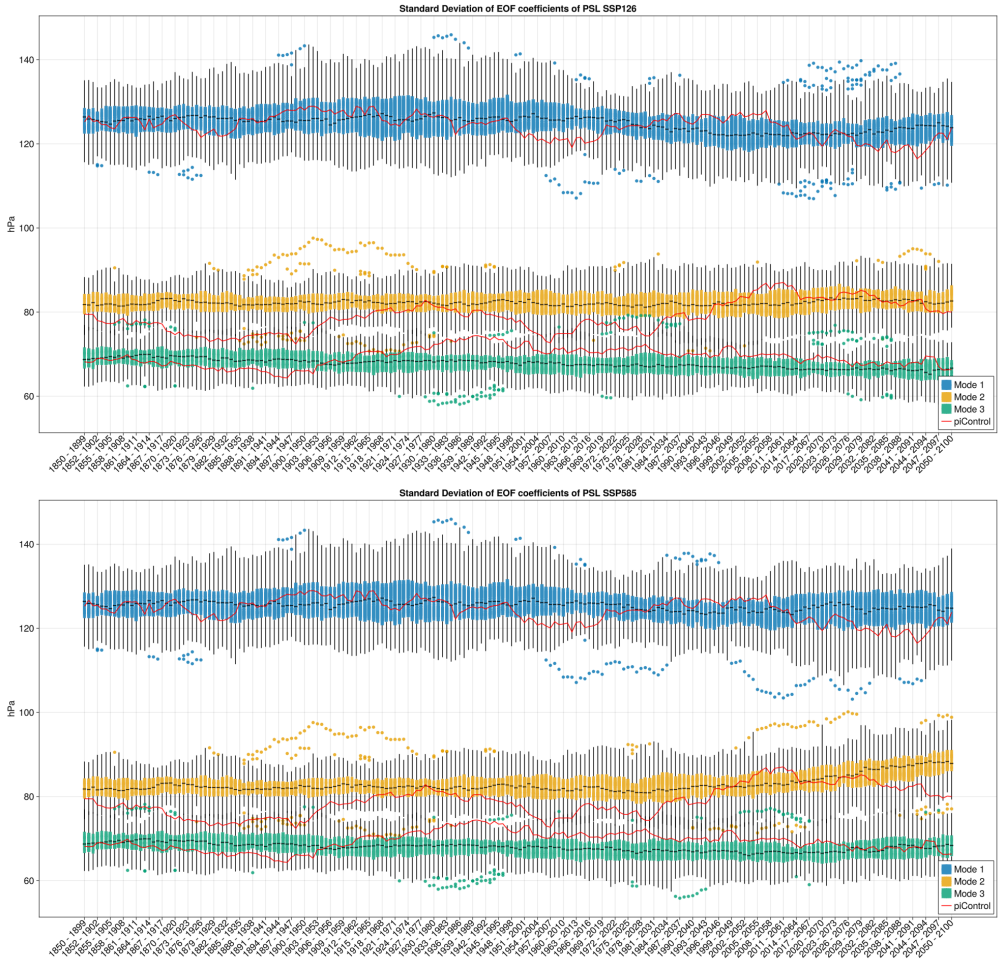


Figure 6.8: Evolution of the standard deviation of PSL EOF coefficients for each scope and simulation member. Top plot shows results for SSP126, bottom plot for SSP585.

a more pronounced increase in the later scopes in SSP585 than in SSP126. This could be explained with the increase in proportional variability of that mode shown in Figure 6.1.

The standard deviation of precipitation EOF coefficients appears to expose the largest differences compared to the piControl simulation and the scenarios. Similarly to the IVT piControl EOF coefficients SD, the precipitation piControl has a dump in EOF coefficient SD in roughly the same timespan of scopes, which may be connected. The SSP126 scenario only shows a moderate increase in mode 1, and no significant change in other modes. On the other hand, SSP585 shows a significant increase in all modes, which is different from the encoded variability shown in Figure 6.3, where only the first mode exposes a steep increase.

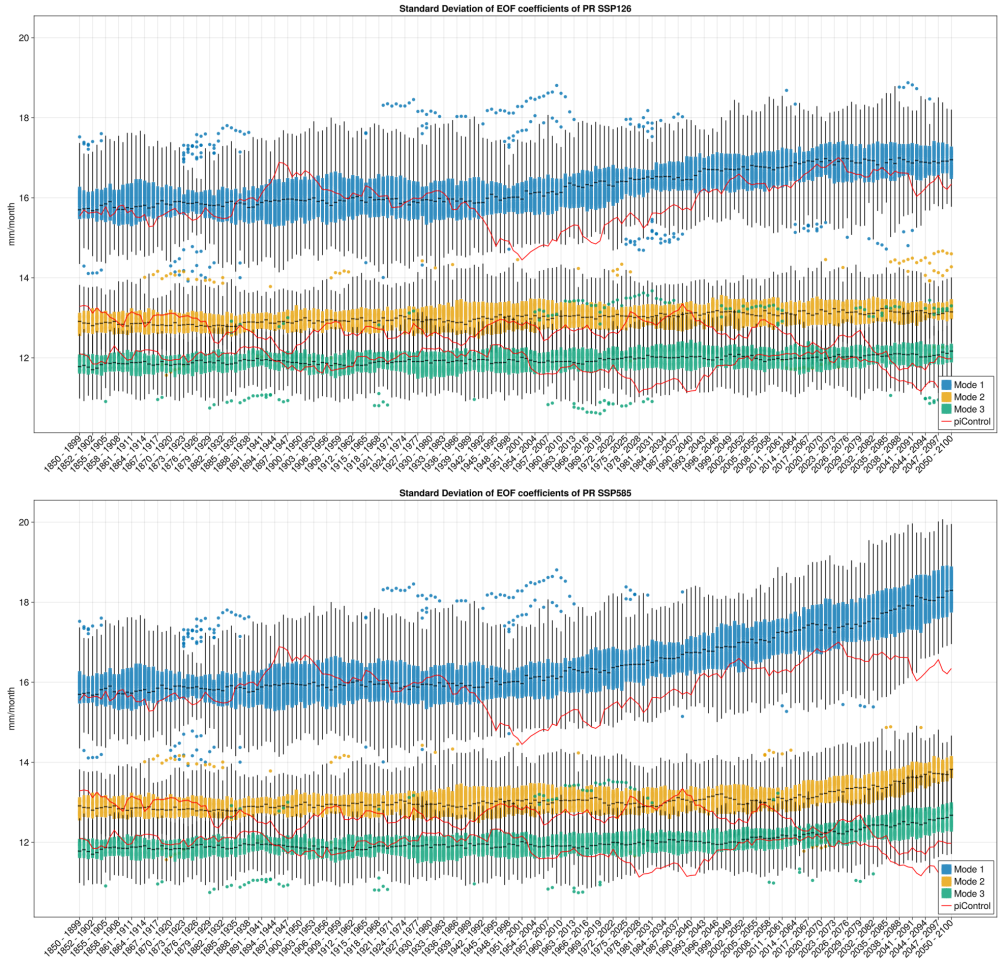


Figure 6.9: Evolution of the standard deviation of Precipitation EOF coefficients for each scope and simulation member. Top plot shows results for SSP126, bottom plot for SSP585.

6.2 Relationships with other Variables

This section explores the correlation relationships between the dominant EOFs and other variables an their EOFs.

6.2.1 Relationships of EOFs

The different EOFs of the variables can easily be compared by measuring the relationship between their temporal EOF coefficients. Although cross-correlation was used to find correlation patterns shifted in time (as explained in Section 5.3), the analysis revealed that all somewhat correlated patterns had the highest correlation with lag zero (≡ the same

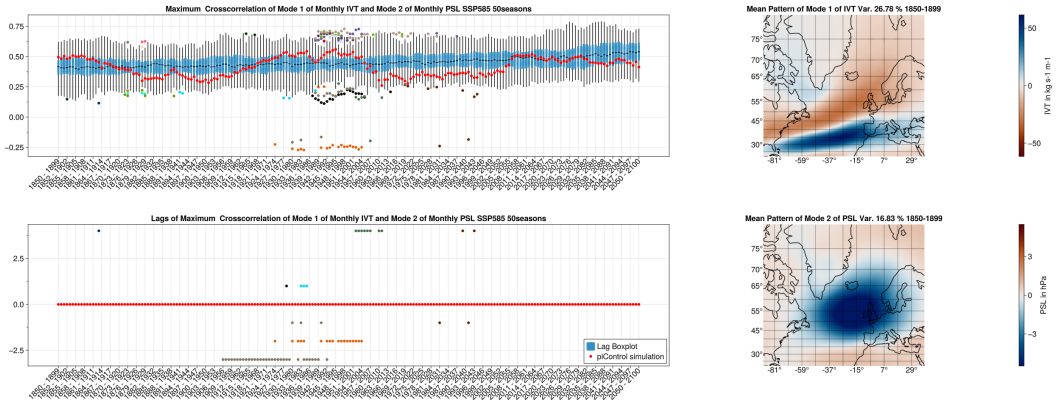


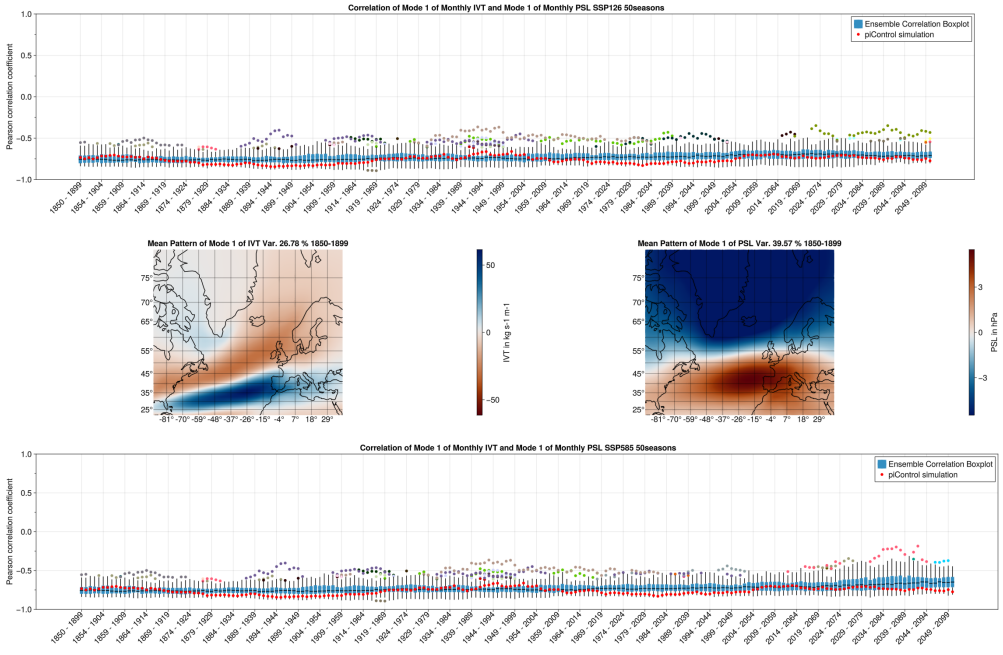
Figure 6.10: Example of cross-correlation analysis of IVT mode 1 and PSL mode 2 in scenario SSP585. The top axis shows a boxplot of the maximum absolute correlation, while the bottom axis shows the associated lag value. Red Line and dots show the piControl run.

month). The few nonzero lag values are only a few outliers, which can be seen in the cross-correlation analysis example of IVT mode 1 and PSL mode 2 in Figure 6.10. Therefore, finding the maximal cross-correlation distorts the actual correlation, which is expected to occur in the same month. That is why this section only displays the correlation that has not shifted. Following the argumentation in the previous section, only relationships between the first two modes of the variables were explored. In addition, only the figures with striking correlations (Pearson correlation coefficient (PCC) > 0.4) are evaluated.

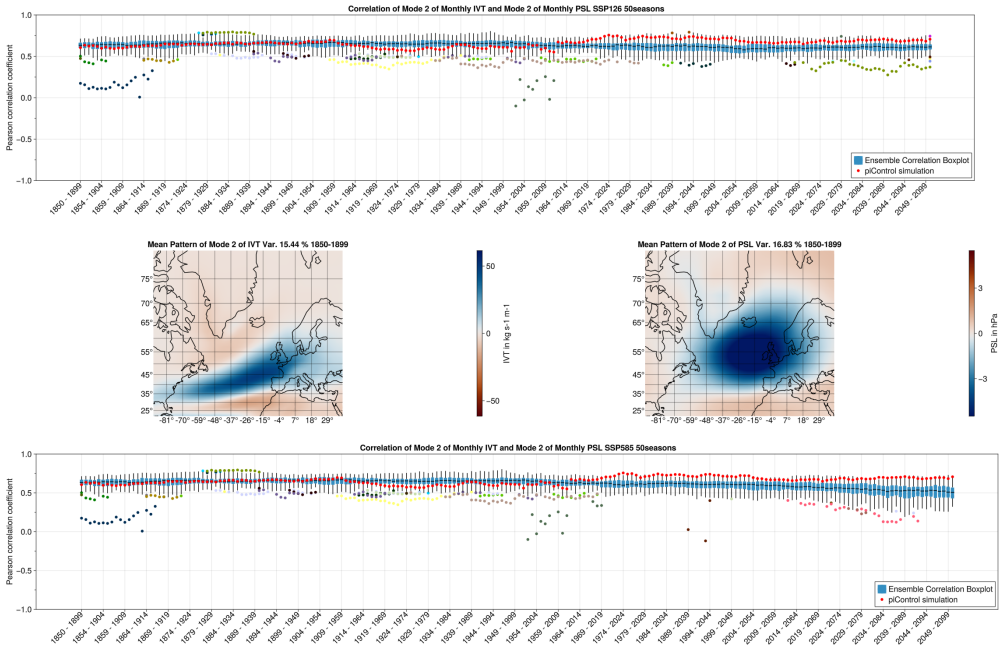
The first relationships to explore are the ones between the indices of the dominant oscillations of the Atlantic and the IVT modes. In general, it is expected that high values of IVT are associated with low pressure areas. Figure 6.11a shows the correlation of the NAO index with the activity of the primary IVT mode. As expected, it shows a strong negative PCC of about -0.75, with quite little variability across the different members. Comparing the different scenarios, it seems that SSP126 stays very similar to the historical simulation. SSP585, on the other hand, seems to be more variable across the members and slightly less correlated in the later years. The first could be the reason for the latter.

The relationships between mode 1 and mode 2 of IVT and PSL and vice versa are pretty similar (Figure 6.10 shows one of them using cross-correlation): Both have the median PCC around 0.5, with the boxplots whiskers reaching around 0.25. Also, both seem to experience a slight increase in median correlation in the late years of SSP585, while SSP126 stays pretty consistent.

Figure 6.11b shows the evolution of the correlation of the index of the EAP with the secondary IVT mode. It shows a slightly weaker relationship than in Figure 6.11a with a



(a) Mode 1 of IVT and Mode 1 of PSL (NAO index)



(b) Mode 2 of IVT and Mode 2 of PSL (EAP index)

Figure 6.11: Correlation of Sea Level Pressure and IVT. Figure 6.11a shows correlations between the temporal patterns of the primary EOFs of PSL and IVT, while Figure 6.11b shows correlation of the secondary modes. Top row in each figure shows SSP126, bottom row SSP585. In the middle are images of the corresponding spatial patterns. The colors of the outliers refer to one specific member associated with that color.

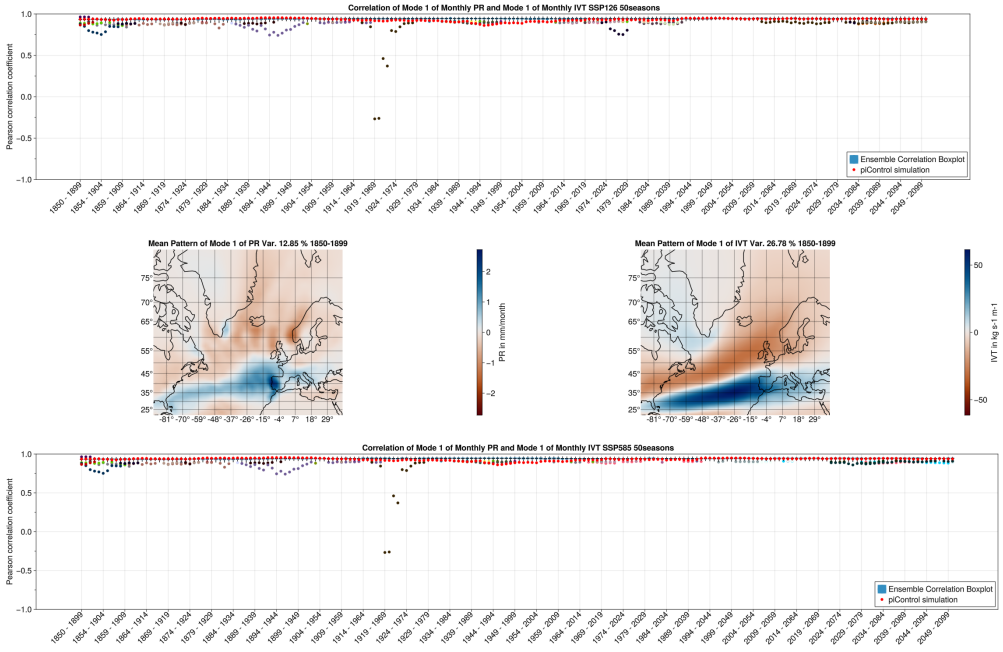
median PCC around 0.6. Again, SSP126 remains similar to the historical simulation and the piControl simulation. But in SSP585, the correlation with the EAP index drops to a median of 0.5, while the variability across members increases.

In Figure 6.12a it is obvious that the primary modes of IVT and precipitation are very closely related. It shows a PCC of about 0.9, with very little variance across the members. In addition, there is barely a difference between SSP126 and SSP585. The secondary modes of IVT and precipitation look quite different: Although Figure 6.12b shows a fairly high median PCC of approximately 0.75, there are a lot of outliers that reach a similar extent in the negative spectrum. Also, the pre-industrial control simulation shows this kind of behavior in the scopes beginning at 1949 to 1974 (with interruptions). The explanation for this is pretty simple: Since this mode is pretty unstable across different members (see previous section), the alignment to the pattern shown in Figure 6.12b did not work very well, which causes the pattern (and following this also the EOF coefficients) to flip, resulting in the value on the other side of the correlation spectrum.

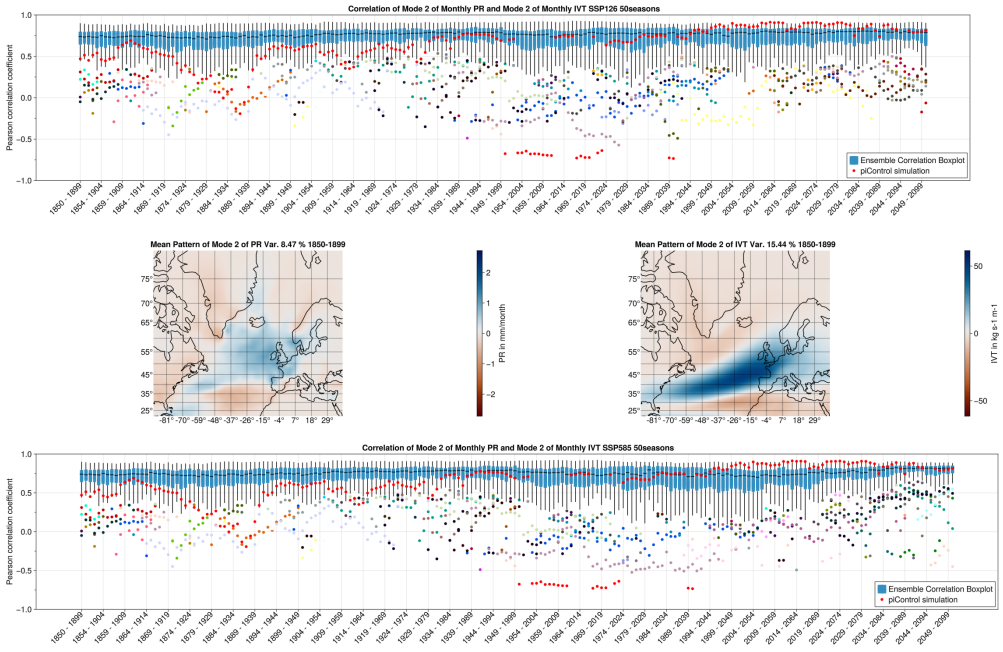
Since the correlation between IVT and NAO/EAP and precipitation and IVT are already known, the analysis of their temporal correlation does not yield surprises. However, Figures 6.13a and 6.13b show their correlation evolution, showing the expected high negative (≈ -0.75) with the NAO index and the expected high positive correlation with the EAP index. Interestingly, the drop in correlation that was visible in Figure 6.11b does not appear in the relationship between precipitation and the EAP. But this could also be due to the amount of outliers (caused by incoherent, misaligned patterns of precipitation EOF2) that warp the results.

6.2.2 Correlation Maps of EOFs and other Fields

Although the relationships of different patterns are interesting, EOFs are notoriously difficult to interpret [DL02; HJS07]. For sea level pressure EOFs this has been solved, since NAO and EAP (and their indices) are established physical modes and were first discovered using traditional measurements and evaluation of weather stations. It was then discovered that the EOFs of said fields also expose the measured patterns and are therefore a useful model to represent these physical modes. This type of analysis does not yet exist for IVT and precipitation in Europe. Since the causality is quite clear (pressure oscillations influence the wind and therefore IVT and IVT influences precipitation), these directions of EOF mode $\rightarrow$ data correlation are explored. The most interesting of such analyzes is the relationship between IVT EOF modes and the precipitation data, since this could help explain what influence certain modes have on the precipitation in Europe. The correlations between sea level pressure EOF modes and IVT data are explored as well.

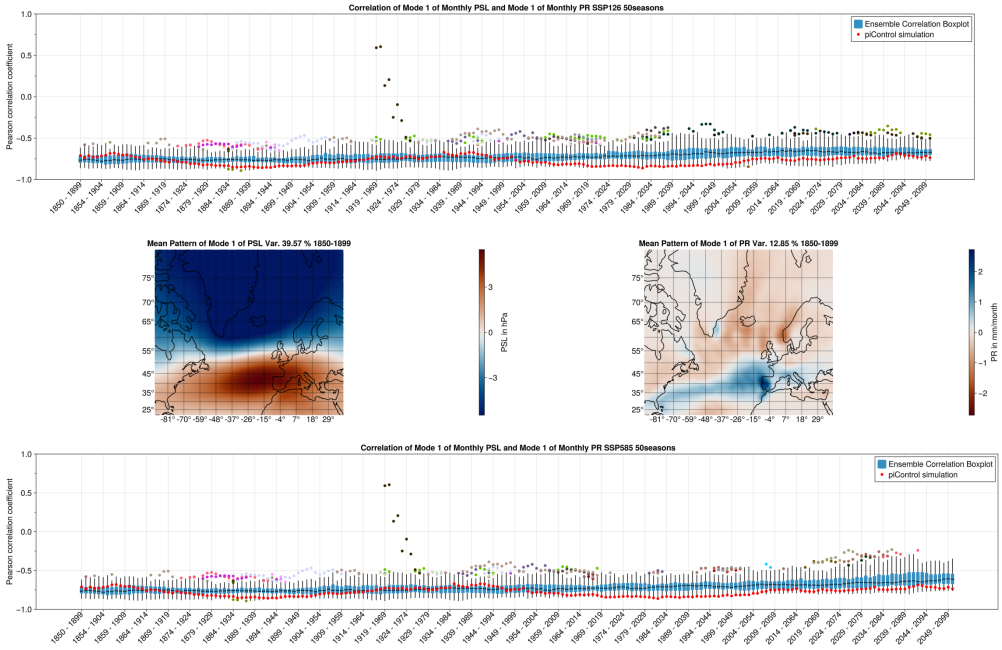


(a) Mode 1 of IVT and Mode 1 of precipitation

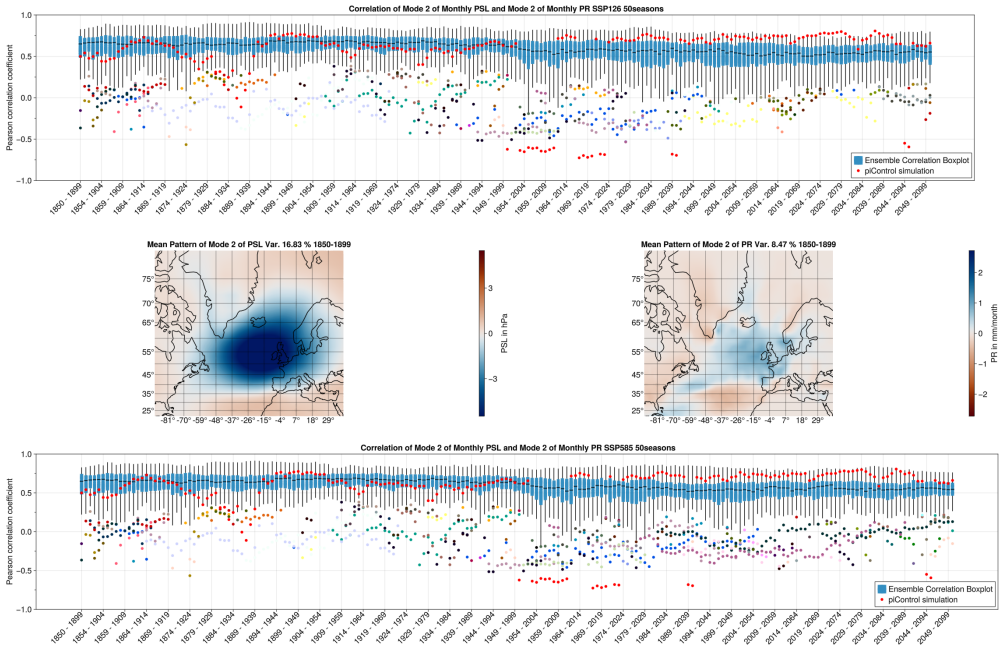


(b) Mode 2 of IVT and Mode 2 of precipitation

Figure 6.12: Correlation of precipitation and IVT. Figure 6.12a shows correlations between the temporal patterns of the primary EOFs of IVT and precipitation, while Figure 6.12b shows correlation of the secondary modes. Top row in each figure shows SSP126, bottom row SSP585. In the middle are images of the corresponding spatial patterns. The colors of the outliers refer to one specific member associated with that color.

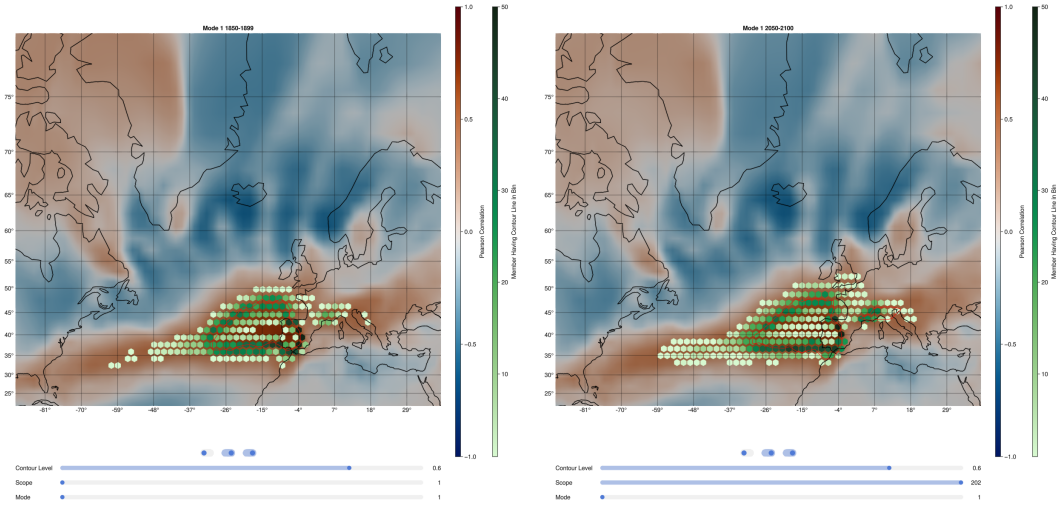


(a) Mode 1 of IVT and Mode 1 of precipitation



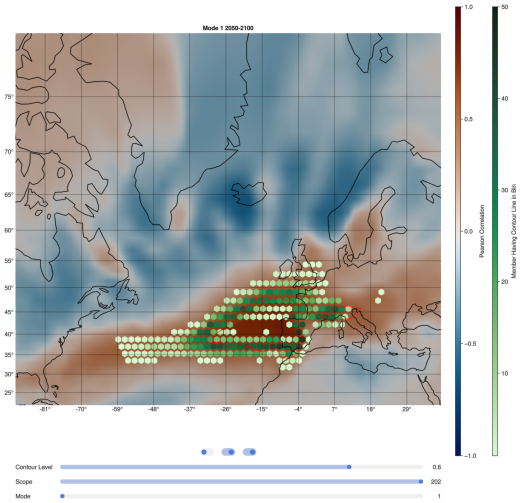
(b) Mode 2 of IVT and Mode 2 of precipitation

Figure 6.13: Correlation of Sea Level Pressure and precipitation. Figure 6.13a shows correlations between the temporal patterns of the primary EOFs of PSL and Precipitation, while Figure 6.13b shows correlation of the secondary modes. Top row in each figure shows SSP126, bottom row SSP585. In the middle are images of the corresponding spatial patterns. The colors of the outliers refer to one specific member associated with that color.



(a) Begin of historical simulation

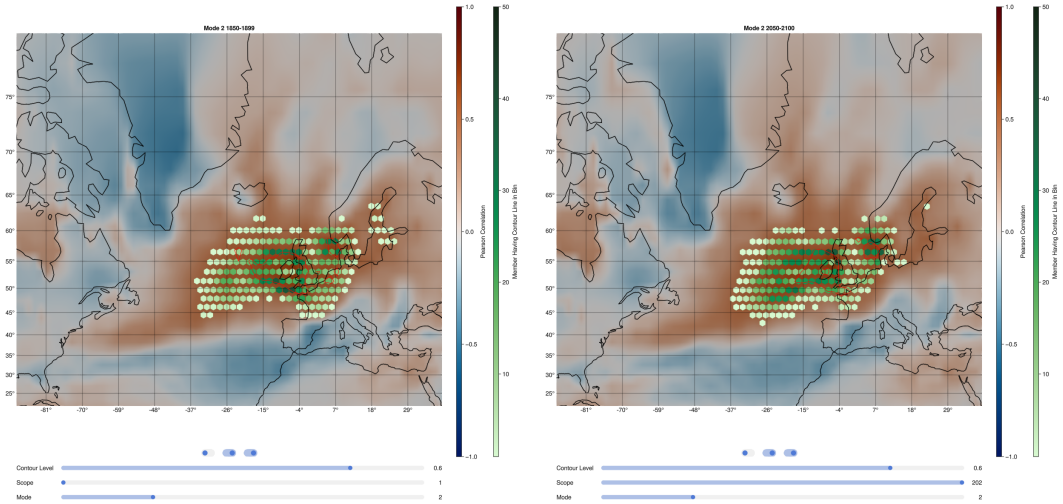
(b) End of SSP126



(c) End of SSP585

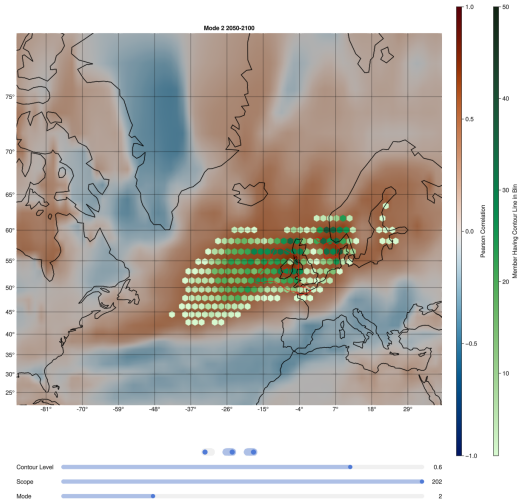
Figure 6.14: Correlation maps of IVT EOF mode 1 and precipitation data of the same scope. Hexbins show the probability of contour lines of 0.6 passing through. The red line indicates the same contour line, but for the preindustrial control simulation.

Figure 6.14 shows the correlation maps of mode 1 of the beginning of the historical simulation, the end of the SSP126 scenario, and the end of SSP585. The isocontours indicate areas with PCC values higher than 0.6, therefore areas where the actual precipitation behaves quite similarly to the dominant IVT pattern. This area is mainly the west coast of the Iberian Peninsula (area of highest correlation), the ocean westward of it, and for some members the Côte d'Azur. Looking at the evolution in time one can see that the area of high correlation do not change that much in SSP126. It seems to stretch a bit further



(a) Begin of historical simulation

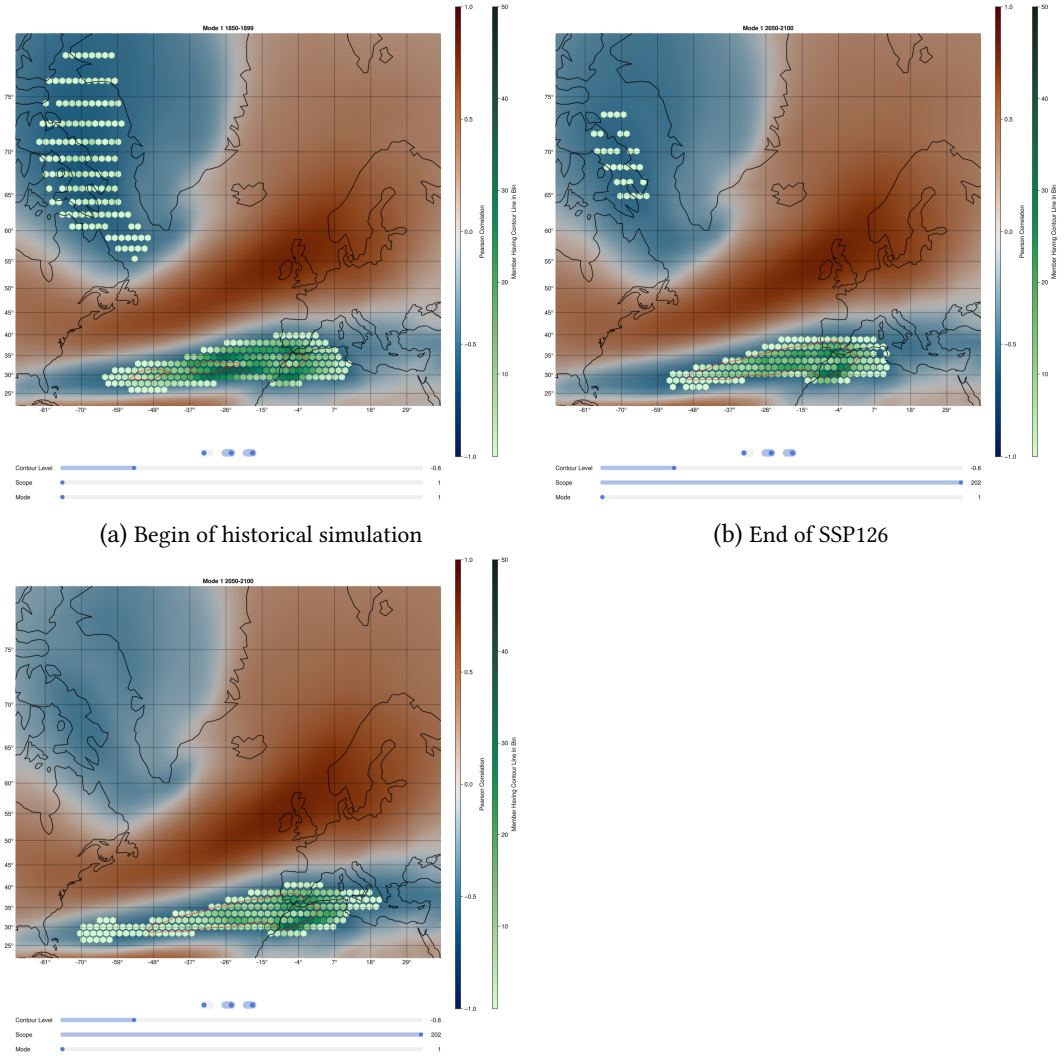
(b) End of SSP126



(c) End of SSP585

Figure 6.15: Correlation maps of IVT EOF mode 2 and precipitation data of the same scope. Hexbins show the probability of contour lines of 0.6 passing through. The red line indicates the same contour line, but for the preindustrial control simulation.

north and more confident in the peak at the Côte d'Azur (darker green of hexbins in that area). Furthermore, the Atlantic coast of France appears to be associated with PCC values > 0.6 in some outliers of the ensemble. SSP585 shows the same changes, but far more pronounced: The outliers even reach the British Islands in the north, and the west coast of France seems to be part of the higher correlation area for most members, similar to the Côte d'Azur.



(a) Begin of historical simulation

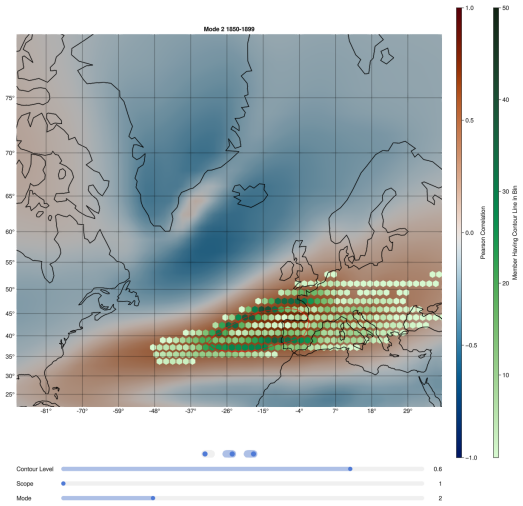
(b) End of SSP126

(c) End of SSP585

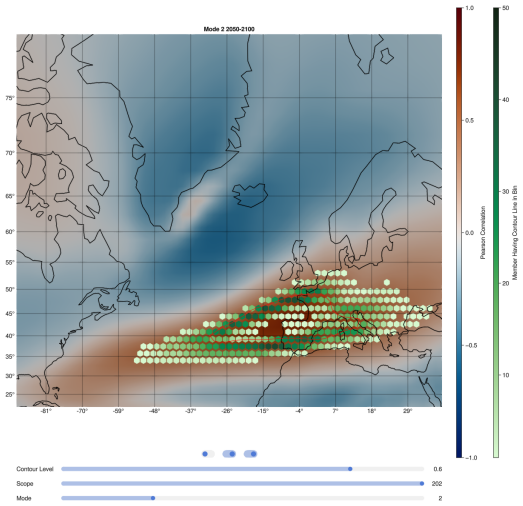
Figure 6.16: Correlation maps of PSL EOF mode 1 (NAO index) and IVT data of the same scope. Hexbins show the probability of contour lines of −0.6 passing through. The red line indicates the same contour line, but for the preindustrial control simulation.

Figure 6.15 shows the same analysis, but for mode 2 of IVT EOF. The area of higher (positive) correlations are here especially the west coasts of the British Islands and Scandinavia, but also the ocean westward of them, the whole west coast of Europe, and the North Sea. Also, there are only a few members indicating that correlation level in the Gulf of Bothnia (between Finland and Sweden), but these seem to be outliers. The end of the SSP126 scenario looks quite similar to the beginning of the historical simulation, except that the contour lines seem to reach less into the European mainland and are less confident

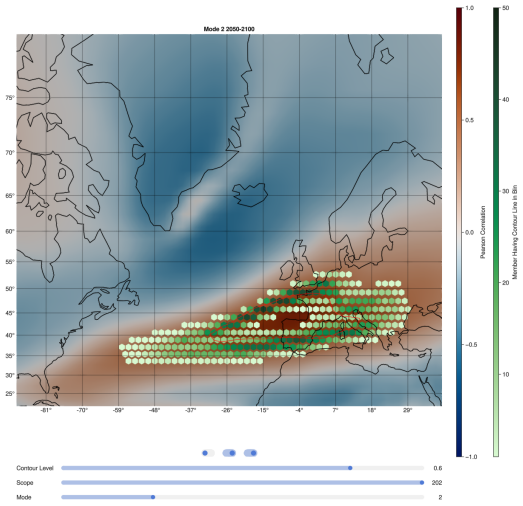
on the west coast of France. This could be due to the increased influence of IVT EOF1 in the same area, which reduces the effect of the secondary mode on that coast.



(a) Begin of historical simulation



(b) End of SSP126



(c) End of SSP585

Figure 6.17: Correlation maps of PSL EOF mode 2 (EAP index) and IVT data of the same scope. Hexbins show the probability of contour lines of −0.6 passing through. The red line indicates the same contour line, but for the preindustrial control simulation.

Figure 6.16 shows the same correlation maps, but with the primary PSL EOF (the NAO index) and IVT data. This has the goal of seeing how the oscillation indices are connected to the actual IVT data, and maybe even tracking a northward shift there. Here, the correlations in the direction of Figures 6.11a and 6.11b, so negative for EOF1 and positive for EOF2. Notable is here the area of positive correlation, which fits the description of the me-

chanics of the NAO given in Section 1.2.2. In general, the hexbins are not as dark green as in the previous figures, indicating that the members do not agree as unanimous. The highest correlation on that level is on the west coast of the Iberian Peninsula and Morocco, and the ocean westward of that area. In the historical simulation, the Labrador Sea also features a few light green hexbins, indicating that only a few members seem to have that kind of correlation between IVT and the NAO index. The changes between the beginning of the historical simulation and the ends of SSP126 and SSP585 are only marginal, no (northward or whatsoever) shift of the really important dark green hexbins. This means that only a few outliers of the members changed, and the contour lines of the largest quantity of members remain pretty much the same.

Figure 6.17 depicts the changes in correlation between the PSL EOF2 coefficients (EAP index) and IVT data. The positive correlation seems to affect all Europe, but especially the European West Coast in France and the Iberian Peninsula. But, similar to the NAO index, there are no structural changes visible between the historical simulation and both future scenarios, only some outliers, while the hexbins of darker green stay in place.

6.3 Discussion

This section discusses the results shown in the previous sections as well as the effectiveness of the introduced hexbin visualization.

6.3.1 Insights about the patterns

The interpretation of applying more complex statistical/mathematical procedures (such as EOFs) to physical data is not trivial and requires the expertise of a domain scientist. While the benefits of analyzing EOFs instead of the original data are great, they are notoriously hard to interpret with regard to the actual atmospheric physics behind them [DL02; HJS07]. The main reasons for this are the forced orthogonality of modes, which does not really exist in the real world [HJS07], and the "hallucination" of modes, which means creating certain modes that do not have an equivalent in the real data/world (which was shown in detail in the work of Dommenget and Latif [DL02]). The reason for said hallucinations is also the forced orthogonality of EOF modes. It is a purely mathematical decomposition based on maximized variability which could be physically interpreted given enough physical indications, such as the NAO index calculated based on weather station data vs. the dominant PSL EOF coefficients. The spatial patterns of EOFs do *not* areas of relative or absolute increase or decrease, but rather show maps of variability patterns which appear together in time. Yet, the importance of the dominant modes of IVT were mentioned and analyzed

in previous work [SRP83; Zou+18]. It could be a great approach to the much needed dimension reduction for the huge amount of turbulent data, focusing only on the most prevalent parts of moisture transport and precipitation. In addition, the results still yield some interesting differences between the historical simulation and both evaluated future scenarios. The focus of this work was a) the generation of EOF patterns of moisture-related variables and b) the visualization of the variability across ensemble members. This section tries to summarize the most important discoveries related to European moisture EOFs, their relationships, and their development in future scenarios.

Starting with the analysis of variance encoded by different modes (Section 6.1.1), the most interesting discoveries are the changing importance of modes in different future scenarios. Most importantly, here is the clearly visible increase in the SD of PSL EOF2 (EAP) in SSP585 by $2-3\%$ and EOF1 of precipitation data by a similar amount. This could be explained by the increase in the area covered by this mode in SSP585, shown in Section 6.1.2. By covering a larger area, this mode can cover a larger portion of the whole datasets' anomaly. In addition, this suggests that a larger area behaves more similarly in time ($\equiv$ can be represented with the same temporal patterns/EOF coefficients).

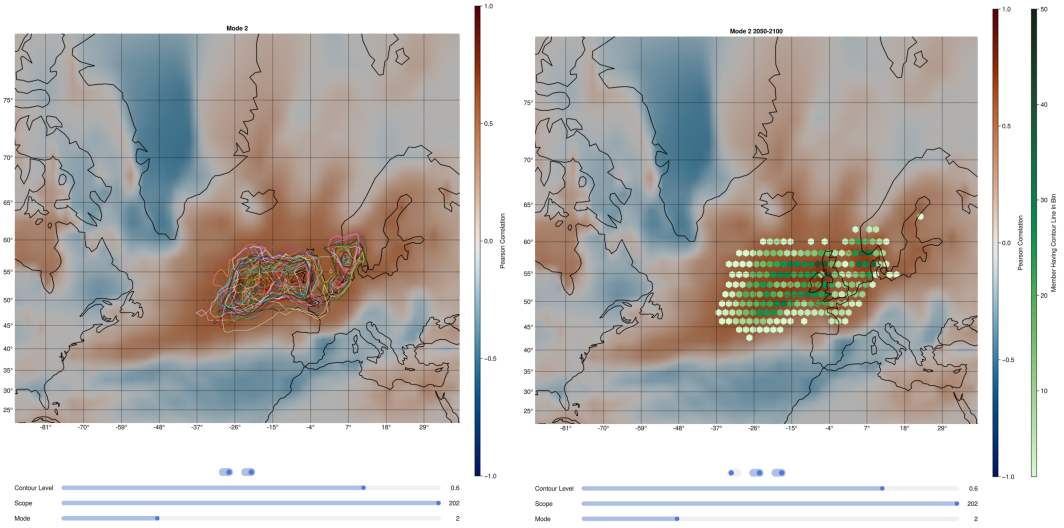
One of the most interesting results of this analysis is the very strong relationship of the primary modes of IVT and precipitation EOFs with a correlation coefficient of about 0.9. This may indicate that EOF1 of IVT is a driver of that particular precipitation mode, as water vapor transport causes precipitation (and not vice versa). The primary mode of PR appears to be especially important for precipitation in the east (coast) of the Iberian Peninsula, which would then also be true for the primary EOF of IVT. This impression is also reinforced by looking at the correlation maps of IVT EOF1 and precipitation data (Figure 6.14), which shows very high correlation values with the temporal pattern of said mode on the east coast of the Iberian Peninsula. The pronounced northward expansion of both PR EOF1 and the correlation between IVT EOF1 and precipitation could also be explained with the northward shift/expansion of the dominant IVT mode, which means that this mode influences a larger region of Europe with more pronounced climate change.

Also interesting is the change in the spatial component of the secondary IVT EOF. Mainly because the consequences of that are not entirely clear. The correlation between IVT EOF2 and precipitation show some changes in SSP585 (less influence on the west coast of Europe), but this could also be due to the increased influence of IVT EOF1. Furthermore, the relationship with the EAP index (Figure 6.11b), appears to drop in the later years of the scenario, in contrast to SSP126.

The analysis of the standard deviation of the different modes had the objective of showing the fluctuations and their evolution. Since the EOF coefficients have a SD of one in

their normal (unscaled) state, they are entirely dependent on the scaling, encoded in the singular values of the respective mode. Most notably is here the striking difference between SSP126 and SSP585 in variance of the dominant IVT EOF modes. Since this is not reflected in the percentage share of variance in these modes, the most likely answer is that the variance of IVT increases in general more in SSP585. A similar pattern can be seen in Figure 6.9, but this can be explained with a combination of an increase in total variance and an increase in the percentage share of PR EOF1. In the end it needs to be stressed that these are just speculations, and the actual interpretation needs to be performed by domain scientists.

6.3.2 Discussion of Hexbin Visualization



(a) Spaghetti plots.

(b) Presented hexbin visualization

Figure 6.18: Comparison of spaghetti plots and the presented contour line visualization using hexbins. The picture is the SSP126 variant shown in Figure 6.15.

This section focuses on the discussion of the presented hexbin visualization of contour lines in ensemble simulations. The goal was to convey the likelihood of a contour line going through a bin in a more direct way, instead of relying on counting (which isn't even possible sometimes). Especially for large numbers of members/contour lines, the real quantity of contour lines passing through an area is quite hard to recognize. Looking at Figure 6.18a, it is quite difficult (even with zooming and counting) to intuitively determine how many contour lines are actually there on the coast of Norway/Sweden. It is easy to recognize very few contour lines, but it gets far harder to distinguish, e.g. between

10 or 25 lines. Although intuitively seeing the exact amount is also not possible using the current hexbin approach, it gives at least an idea about the percentage and can be compared to other regions (e.g., the density at the coast of Norway/Sweden is similar to the density near Ireland). In the authors' opinion, it is not really possible to compare in spaghetti plots. Given the fact that the trend of ensemble members tends to be more than less, spaghetti plots seem to be less capable of displaying an ensemble's variability.

A weakness of the approach is the distortion introduced by the map projection, which results in the space between the bins in the far north of the map, which may result in confusion. A possible way of fixing this error could be to also distort the polygons at each position, similar to the Tissot's spheres shown in Section 2.1.2.

7 Conclusions and Future Work

7.1 Conclusions

This thesis presents the first ensemble-scale evaluation of Empirical Orthogonal Functions (EOFs) for moisture-related variables across Europe, encompassing the two extremes of CMIP6 future scenarios. A sliding window approach was used in the EOF analysis to evaluate the evolution of the patterns over time. The primary objective of this pattern analysis was to gain visual insights into the structural nature of these recurring oscillations of moisture transport and their consequences, which are typically challenging to track. Furthermore, the EOF analysis has the advantage of reducing the complexity of the dataset, making it easier to realize the visualization of the ensembles' multiple members. This work provides the first known computation of IVT and precipitation EOF patterns for Europe and the northern Atlantic. The relationships between these patterns and other variables were explored to enhance the understanding of their implications. This thesis focuses on two major oscillations in Europe, the North Atlantic Oscillation (NAO) and the East Atlantic Pattern (EAP), which are established results of EOF analysis of sea level pressure data and primary drivers of winter winds and weather patterns. Additionally, the most critical result of moisture transport, precipitation, was analyzed to explore the consequences and significance of specific modes of moisture transport. To illustrate member variability, contour lines were employed. Moreover, the thesis introduced a method for displaying multiple contour lines using hexbins, enhancing the clarity and interpretability of the data compared to the conventional method (i.e., spaghetti plots).

The analysis focused on stationary water vapor transport, represented by monthly means, and its relationship with monthly precipitation and sea level pressure, which reflect the dominant oscillations, NAO and EAP. The results identified two dominant modes of water vapor transport which demonstrated considerable stability across different members and time periods. In the analysis of precipitation EOFs, the primary mode appeared to be quite stable, while the secondary mode appeared questionable, with the remaining modes exhibiting degeneracy. Relationships between these patterns were explored using two correlation-based methods: one involved comparing the modes of sea level pressure

(PSL), IVT, and precipitation EOFs, while the other involved generating correlation maps between specific modes and the corresponding data. The first method revealed strong correlations between the dominant modes of PSL and IVT, as well as between these modes and precipitation. However, the correlations were somewhat weaker in the secondary EOFs, but still exist. A particularly strong relationship was observed between the primary modes of IVT and precipitation, which could be relevant to understand precipitation variability in the Iberian Peninsula. Conversely, the connection between the secondary EOFs of IVT and precipitation was weaker, possibly due to the instability of the secondary precipitation EOF, which raises questions about its interpretative usefulness. The second method indicated that the area of precipitation strongly correlated with the primary mode extends northward, with this pattern becoming more pronounced under scenarios of more severe climate change. Furthermore, the secondary mode of IVT, and consequently the precipitation data, appeared to have less influence on the western coast of Europe in more pronounced climate change.

Reflecting on this thesis, the choice to utilize the Julia language and its related framework proved to be somewhat questionable. Although Julia has gained significant traction within the geo-scientific community, it has not yet achieved the same level of maturity as Python. While Julia excels in implementing mathematical concepts, Python offers a more extensive array of libraries that facilitate various tasks. The Makie framework, though excellent for visualization, suffers from inadequate documentation, making adjustments challenging. Additionally, the GeoMakie library, responsible for the map projections of the data, is not sufficiently mature for the applications required in this thesis, presenting several limitations, such as issues with projections and boundaries. The proposed hexbin-based visualizatio despite ability to mitigate some issues of spaghetti plots, seems to not be better than any already existing uncertain contour line visualization techniques. Therefor, an useable implementation of them could be very useful, since the biggest problem of the already existing approaches is the lack of implementation in all-purpose visualization toolkits. While the results suggest changes potentially related to climate change, the interpretation of the EOF analysis remains complex. Consequently, this type of analysis requires validation and acceptance from meteorologists and climate scientists to ensure its credibility and applicability.

7.2 Future Work

First, anything that helps to understand the consequences of certain EOF modes could solidify this kind of analysis of water vapor transport. For example, applying EOF analysis

to reanalysis data of similar variables could offer further insights into the significance of identified modes. Or by e.g. analyzing how much of Spain's precipitation can be explained by the dominant precipitation mode by comparing the reconstruction of this mode with actual data, could provide valuable insights. Also, incorporating a broader range of statistical methods, such as Spearman correlation and regression analysis, could enhance the comparative analysis of different visualizations. These methods can also be easily utilized in the presented visual analysis techniques in this Thesis, like correlation boxplots of temporal patterns and correlation maps. Implementing the established visualization techniques for contour lines or scalar fields, such as contour boxplots, could improve the representation of member variability. So for example the contour boxplot approach presented in related work can be combined with clustering algorithms to display the uncertainties in level crossing probabilities. This would aid in the clearer visualization of differences and patterns within the data. As pointed out in Chapter 5, for this at first the far higher timely resolution of 6-hourly data was computed, which can also be analyzed to get insights into the transient components of moisture transport, which could reuse much of the work done for this thesis. This could also be used to be linked to atmospheric rivers, which seems to be an area of great interest in the atmospheric science community. Furthermore, exploring alternative pattern analysis methods, such as Self-Organizing Maps (SOMs, see Teale and Robinson [TR20]), might also be beneficial. However, these methods often face challenges related to interpretability, which need to be addressed. Additionally, investigating the topological structure of the modes and employing approaches similar to those suggested Vietinghoff et al. [Vie+21a] could provide new perspectives and enhance the overall analysis. By pursuing these directions, future research can build on the findings of this thesis, contributing to a more comprehensive understanding of atmospheric patterns of moisture transport and their implications.

Acronyms

CMIP Coupled Model Intercomparison Project

EAP East Atlantic Pattern

ENSO El Niño Southern Oscillation

EOF Empirical Orthogonal Functions

ERF Effective Radiative Forcing

GCM Global Coupled Model

GHG Greenhouse Gasses

HPC high performance computing

IPCC Intergovernmental Panel on Climate Change

IVT Integrated Water Vapor Transport

IWV Vertically Integrated Water Vapor

MIP Model Intercomparison Project

MPI GE CMIP6 Max Planck Institute Grand Ensemble CMIP6

NAO North Atlantic Oscillation

PCA Principal Component Analysis

PCC Pearson correlation coefficient

PDF Probability Density Function

PSL Sea Level Pressure

RCP Representative Concentration Pathways

SD standard deviation

SOM Self Organizing Map

SSP Shared Socioeconomic Pathways

SST Sea Surface Temperature

SVD Singular Value Decomposition

Bibliography

[24] *JuliaIO/JLD2.jl*. 26, 2024. URL: https://github.com/JuliaIO/JLD2.jl (visited on 07/01/2024).

[Aya+22] O. O. Ayantobo, J. Wei, B. Kang, and G. Wang. "Integrated moisture transport variability over China: patterns, impacts, and relationship with El Nino–Southern Oscillation (ENSO)". *Theoretical and Applied Climatology* 147:3, 2022, pp. 985–1002. ISSN: 0177-798X, 1434-4483. DOI: 10.1007/s00704-021-03864-x.

[Bao+06] J.-W. Bao, S. A. Michelson, P. J. Neiman, F. M. Ralph, and J. M. Wilczak. "Interpretation of Enhanced Integrated Water Vapor Bands Associated with Extratropical Cyclones: Their Formation and Connection to Tropical Moisture". *Monthly Weather Review* 134:4, 1, 2006, pp. 1063–1080. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/MWR3123.1.

[Bar24] A. Barth. *NCDatasets.jl: a Julia package for manipulating netCDF data sets*. 2024. DOI: 10.21105/joss.06504.

[Ben88] Y. Benjamini. "Opening the Box of a Boxplot". *The American Statistician* 42:4, 1, 1988, pp. 257–262. ISSN: 0003-1305. DOI: 10.1080/00031305.1988.10475580.

[Bez+17] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. "Julia: A Fresh Approach to Numerical Computing". *SIAM Review* 59:1, 2017, pp. 65–98. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/141000671.

[BK] M. Böttinger and D. D. Kasang. *The SSP Scenarios*. DKRZ. URL: https://www.dkrz.de/en/communication/climate-simulations/cmip6-en/the-ssp-scenarios (visited on 06/12/2024).

[BKS04] U. D. Bordoloi, D. L. Kao, and H.-W. Shen. "Visualization techniques for spatial probability density function data". *Data Science Journal* 3, 2004, pp. 153–162. ISSN: 1683-1470. DOI: 10.2481/dsj.3.153.

[Bro04] R. Brown. "Animated visual vibrations as an uncertainty visualisation technique". In: *Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*. GRAPHITE '04. Association for Computing Machinery, New York, NY, USA, 15, 2004, pp. 84–89. ISBN: 978-1-58113-883-2. DOI: 10.1145/988834.988849.

[Buc] G. Buckley. *Choosing good chunk sizes in Dask*. URL: https://blog.dask.org/2021/11/02/choosing-dask-chunk-sizes (visited on 06/25/2024).

[CMW16] L. Comas-Bru, F. McDermott, and M. Werner. "The effect of the East Atlantic pattern on the precipitation δ18O-NAO relationship in Europe". *Climate Dynamics* 47:7, 2016, pp. 2059–2069. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-015-2950-1.

[Con+11] A. Coninx, G.-P. Bonneau, J. Droulez, and G. Thibault. "Visualization of uncertain scalar data fields using color scales and perceptually adapted noise". In: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*. APGV '11: ACM Symposium on Applied Perception in Graphics & Visualization 2011. ACM, Toulouse France, 27, 2011, pp. 59–66. ISBN: 978-1-4503-0889-2. DOI: 10.1145/2077451.2077462.

[COW92] D. B. Carr, A. R. Olsen, and D. White. "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data". *Cartography and Geographic Information Systems* 19:4, 1992, pp. 228–236. ISSN: 1050-9844. DOI: 10.1559/152304092783721231.

[DK21] S. Danisch and J. Krumbiegel. "Makie.jl: Flexible high-performance data visualization for Julia". *Journal of Open Source Software* 6:65, 1, 2021, p. 3349. ISSN: 2475-9066. DOI: 10.21105/joss.03349.

[DL02] D. Dommenget and M. Latif. "A Cautionary Note on the Interpretation of EOFs". *Journal of Climate* 15:2, 15, 2002, pp. 216–225. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/1520-0442(2002)015<0216:ACNOTI>2.0.CO;2.

[EBM16] J. Eiras-Barca, S. Brands, and G. Miguez-Macho. "Seasonal variations in North Atlantic atmospheric river activity and associations with anomalous precipitation over the Iberian Atlantic Margin". *Journal of Geophysical Research: Atmospheres* 121:2, 2016, pp. 931–948. ISSN: 2169-8996. DOI: 10.1002/2015JD023379.

[Eck09] S. Eckermann. "Hybrid σ–p Coordinate Choices for a Global Model". *Monthly Weather Review* 137:1, 1, 2009, pp. 224–245. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/2008MWR2537.1.

[Eyr+16] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". *Geoscientific Model Development* 9:5, 26, 2016, pp. 1937–1958. ISSN: 1991-9603. DOI: 10.5194/gmd-9-1937-2016.

[Fol+11] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson. "An overview of the HDF5 technology suite and its applications". In: *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. EDBT/ICDT '11: EDBT/ICDT '11 joint conference. ACM, Uppsala Sweden, 25, 2011, pp. 36–47. ISBN: 978-1-4503-0614-0. DOI: 10.1145/1966895.1966900.

[Foo56] E. Foote. "Circumstances affecting the heat of the sun's rays". *Am. J. Sci. Arts* 22:66, 1856, pp. 383–384.

[Fou24] J. Fourier. "Remarques générales sur les températures du globe terrestre et des espaces planétaires". In: *Annales de Chemie et de Physique*. Vol. 27. 1824, pp. 136–167.

[FSZ03] J. Fernández, J. Sáenz, and E. Zorita. "Analysis of wintertime atmospheric moisture transport and its variability over southern Europe in the NCEP Reanalyses". *Climate Research* 23, 2003, pp. 195–215. ISSN: 0936-577X, 1616-1572. DOI: 10.3354/cr023195.

[Gao+20] K. Gao, G. Mei, F. Piccialli, S. Cuomo, J. Tu, and Z. Huo. "Julia language in machine learning: Algorithms, applications, and open issues". *Computer Science Review* 37, 1, 2020, p. 100254. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2020.100254.

[Gha16] E. Ghaderpour. "Map Projection". *Journal of Applied Geodesy* 10:3, 1, 2016, pp. 197–209. ISSN: 1862-9024, 1862-9016. DOI: 10.1515/jag-2015-0033. arXiv: 1412.7690[physics].

[Gim+14] L. Gimeno, R. Nieto, M. Vázquez, and D. Lavers. "Atmospheric rivers: a mini-review". *Frontiers in Earth Science* 2, 2014. ISSN: 2296-6463.

[Gui+18] K. Guirguis, A. Gershunov, R. E. S. Clemesha, T. Shulgina, A. C. Subramanian, and F. M. Ralph. "Circulation Drivers of Atmospheric Rivers at the North American West Coast". *Geophysical Research Letters* 45:22, 2018, pp. 12, 576–12, 584. ISSN: 1944-8007. DOI: 10.1029/2018GL079249.

[Han] A. Hannachi. "A Primer for EOF Analysis of Climate Data".

[HH17] S. Hoyer and J. Hamman. "xarray: N-D labeled Arrays and Datasets in Python". *Journal of Open Research Software* 5:1, 5, 2017, pp. 10–10. ISSN: 2049-9647. DOI: 10.5334/jors.148.

[HJS07] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson. "Empirical orthogonal functions and related techniques in atmospheric science: A review". *International Journal of Climatology* 27:9, 2007, pp. 1119–1152. ISSN: 08998418, 10970088. DOI: 10.1002/joc.1499.

[Hur+03] J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. "An overview of the North Atlantic Oscillation". In: *Geophysical Monograph Series*. Ed. by J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. Vol. 134. American Geophysical Union, Washington, D. C., 2003, pp. 1–35. ISBN: 978-0-87590-994-3. DOI: 10.1029/134GM01.

[Int23] Intergovernmental Panel On Climate Change (Ipcc). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press, 6, 2023. ISBN: 978-1-00-915789-6. DOI: 10.1017/9781009157896.

[Joh04] C. Johnson. "Top scientific visualization research problems". *IEEE Computer Graphics and Applications* 24:4, 2004, pp. 13–17. ISSN: 1558-1756. DOI: 10.1109/MCG.2004.20.

[KA15] H.-M. Kim and M. A. Alexander. "ENSO's Modulation of Water Vapor Transport over the Pacific–North American Region". *Journal of Climate* 28:9, 1, 2015, pp. 3846–3856. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-14-00725.1.

[Kam+21] A. Kamal, P. Dhakal, A. Y. Javaid, V. K. Devabhaktuni, D. Kaur, J. Zaientz, and R. Marinier. "Recent advances and challenges in uncertainty visualization: a survey". *Journal of Visualization* 24:5, 2021, pp. 861–890. ISSN: 1343-8875, 1875-8975. DOI: 10.1007/s12650-021-00755-1.

[Kao+02] D. Kao, A. Luo, J. Dungan, and A. Pang. "Visualizing spatially varying distribution data". In: *Proceedings Sixth International Conference on Information Visualisation*. Proceedings Sixth International Conference on Information Visualisation. 2002, pp. 219–225. DOI: 10.1109/IV.2002.1028780.

[KBR22] M. Kaltenbacher, V. Badeli, and A. Reinbacher-Köstinger. "Nonconforming finite element formulation for the simulation of impedance cardiography". *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields* 36, 16, 2022. DOI: 10.1002/jnm.3063.

[Lan+24] F. Lan, B. Gamelin, L. Yan, J. Wang, B. Wang, and H. Guo. "Topological Characterization and Uncertainty Visualization of Atmospheric Rivers". *Computer Graphics Forum* 43:3, 2024, e15084. ISSN: 1467-8659. DOI: 10.1111/cgf.15084.

[Lee+24] H. Lee, K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. Thorne, C. Trisos, J. Romero, P. Aldunce, and A. C. Ruane. "Climate change 2023 synthesis report summary for policymakers". *CLIMATE CHANGE 2023 Synthesis Report: Summary for Policymakers*, 2024.

[Lob+20] D. Lobelle, C. Beaulieu, V. Livina, F. Sévellec, and E. Frajka-Williams. "Detectability of an AMOC Decline in Current and Projected Climate Changes". *Geophysical Research Letters* 47:20, 2020, e2020GL089974. ISSN: 1944-8007. DOI: 10.1029/2020GL089974.

[LZ12] X. Li and W. Zhou. "Quasi-4-Yr Coupling between El Niño–Southern Oscillation and Water Vapor Transport over East Asia–WNP". *Journal of Climate* 25:17, 1, 2012, pp. 5879–5891. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-11-00433.1.

[Ma+18] Y. Ma, M. Lu, H. Chen, M. Pan, and Y. Hong. "Atmospheric moisture transport versus precipitation across the Tibetan Plateau: A mini-review and current challenges". *Atmospheric Research* 209, 1, 2018, pp. 50–58. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2018.03.015.

[Mac+12] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. "Visual Semiotics & Uncertainty Visualization: An Empirical Study". *IEEE Transactions on Visualization and Computer Graphics* 18:12, 2012, pp. 2496–2505. ISSN: 1941-0506. DOI: 10.1109/TVCG.2012.279.

[Mah+19] N. Maher, S. Milinski, L. Suarez-Gutierrez, M. Botzet, M. Dobrynin, L. Kornblueh, J. Kröger, Y. Takano, R. Ghosh, C. Hedemann, C. Li, H. Li, E. Manzini, D. Notz, D. Putrasahan, L. Boysen, M. Claussen, T. Ilyina, D. Olonscheck, T. Rad-

datz, B. Stevens, and J. Marotzke. "The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability". *Journal of Advances in Modeling Earth Systems* 11:7, 2019, pp. 2050–2069. ISSN: 1942-2466, 1942-2466. DOI: 10.1029/2019MS001639.

[MML21] N. Maher, S. Milinski, and R. Ludwig. "Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble". *Earth System Dynamics* 12:2, 22, 2021, pp. 401–418. ISSN: 2190-4987. DOI: 10.5194/esd-12-401-2021.

[Nei+08] P. J. Neiman, F. M. Ralph, G. A. Wick, J. D. Lundquist, and M. D. Dettinger. "Meteorological Characteristics and Overland Precipitation Impacts of Atmospheric Rivers Affecting the West Coast of North America Based on Eight Years of SSM/I Satellite Observations". *Journal of Hydrometeorology* 9:1, 1, 2008, pp. 22–47. ISSN: 1525-7541, 1525-755X. DOI: 10.1175/2007JHM855.1.

[NOA] NOAA. *What's the difference between climate and weather? | National Oceanic and Atmospheric Administration*. URL: https://www.noaa.gov/explainers/what-s-difference-between-climate-and-weather (visited on 06/19/2024).

[Nor+82] G. R. North, T. L. Bell, R. F. Cahalan, and F. J. Moeng. "Sampling Errors in the Estimation of Empirical Orthogonal Functions". *Monthly Weather Review* 110:7, 1, 1982, pp. 699–706. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.

[Olo+23] D. Olonscheck, L. Suarez-Gutierrez, S. Milinski, G. Beobide-Arsuaga, J. Baehr, F. Fröb, L. Hellmich, T. Ilyina, C. Kadow, D. Krieger, H. Li, J. Marotzke, É. Plésiat, M. Schupfner, F. Wachsmann, K.-H. Wieners, and S. Brune. *The new Max Planck Institute Grand Ensemble with CMIP6 forcing and high-frequency model output*. preprint. Preprints, 4, 2023. DOI: 10.22541/essoar.168319746.64037439/v1.

[ONe+16] B. C. O'Neill, C. Tebaldi, D. P. Van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi, and B. M. Sanderson. "The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6". *Geoscientific Model Development* 9:9, 28, 2016, pp. 3461–3482. ISSN: 1991-9603. DOI: 10.5194/gmd-9-3461-2016.

[Pöt15] K. Pöthkow. "Modeling, Quantification and Visualization of Probabilistic Features in Fields with Uncertainties", 2015.

[PPH13] K. Poethkow, C. Petz, and H.-C. Hege. "Approximate Level-Crossing Probabilities for Interactive Visualization of Uncertain Isocontours". *International Journal for Uncertainty Quantification* 3:2, 2013, pp. 101–117. ISSN: 2152-5080. DOI: 10.1615/Int.J.UncertaintyQuantification.2012003958.

[PWH11] K. Pöthkow, B. Weber, and H.-C. Hege. "Probabilistic Marching Cubes". *Computer Graphics Forum* 30:3, 2011, pp. 931–940. ISSN: 1467-8659. DOI: 10.1111/j.1467-8659.2011.01942.x.

[Ram+16] A. M. Ramos, R. Nieto, R. Tomé, L. Gimeno, R. M. Trigo, M. L. R. Liberato, and D. A. Lavers. "Atmospheric rivers moisture sources from a Lagrangian perspective". *Earth System Dynamics* 7:2, 22, 2016, pp. 371–384. ISSN: 2190-4987. DOI: 10.5194/esd-7-371-2016.

[RD90] R. Rew and G. Davis. "NetCDF: an interface for scientific data access". *IEEE Computer Graphics and Applications* 10:4, 1990, pp. 76–82. ISSN: 0272-1716. DOI: 10.1109/38.56302.

[Ria+17] K. Riahi, D. P. Van Vuuren, E. Kriegler, J. Edmonds, B. C. O'Neill, S. Fujimori, N. Bauer, K. Calvin, R. Dellink, O. Fricko, W. Lutz, A. Popp, J. C. Cuaresma, S. Kc, M. Leimbach, L. Jiang, T. Kram, S. Rao, J. Emmerling, K. Ebi, T. Hasegawa, P. Havlik, F. Humpenöder, L. A. Da Silva, S. Smith, E. Stehfest, V. Bosetti, J. Eom, D. Gernaat, T. Masui, J. Rogelj, J. Strefler, L. Drouet, V. Krey, G. Luderer, M. Harmsen, K. Takahashi, L. Baumstark, J. C. Doelman, M. Kainuma, Z. Klimont, G. Marangoni, H. Lotze-Campen, M. Obersteiner, A. Tabeau, and M. Tavoni. "The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview". *Global Environmental Change* 42, 2017, pp. 153–168. ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2016.05.009.

[Rip+19] W. J. Ripple, C. Wolf, T. M. Newsome, P. Barnard, and W. R. Moomaw. "World Scientists' Warning of a Climate Emergency". *BioScience*, 5, 2019, biz088. ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/biz088.

[Roc+15] M. Rocklin et al. "Dask: Parallel computation with blocked algorithms and task scheduling." In: *SciPy*. 2015, pp. 126–132.

[San+10] J. Sanyal, Song Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. "Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty". *IEEE Transactions on Visualization and Computer Graphics* 16:6, 2010, pp. 1421–1430. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.181.

[Sch24] U. Schulzweida. *CDO - Climate Data Operators*. Accessed: 2024-06-06. 2024.

[SE90] P. Schluessel and W. J. Emery. "Atmospheric water vapour over oceans from SSM/I measurements". *International Journal of Remote Sensing* 11:5, 1990, pp. 753–766. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431169008955055.

[Sea+20] R. Seager, H. Liu, Y. Kushnir, T. J. Osborn, I. R. Simpson, C. R. Kelley, and J. Nakamura. "Mechanisms of Winter Precipitation Variability in the European–Mediterranean Region Associated with the North Atlantic Oscillation". *Journal of Climate* 33:16, 15, 2020, pp. 7179–7196. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-20-0011.1.

[SLP19] Z. Song, M. Latif, and W. Park. "East Atlantic Pattern Drives Multidecadal Atlantic Meridional Overturning Circulation Variability During the Last Glacial Maximum". *Geophysical Research Letters* 46:19, 16, 2019, pp. 10865–10873. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2019GL082960.

[Sou+20] P. M. Sousa, A. M. Ramos, C. C. Raible, M. Messmer, R. Tomé, J. G. Pinto, and R. M. Trigo. "North Atlantic Integrated Water Vapor Transport—From 850 to 2100 CE: Impacts on Western European Rainfall". *Journal of Climate* 33:1, 1, 2020, pp. 263–279. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-19-0348.1.

[SRP83] D. A. Salstein, R. D. Rosen, and J. P. Peixoto. "Modes of Variability in Annual Hemispheric Water Vapor and Transport Fields". *Journal of the Atmospheric Sciences* 40:3, 1983, pp. 788–804. ISSN: 0022-4928, 1520-0469. DOI: 10.1175/1520-0469(1983)040<0788:MOVIAH>2.0.CO;2.

[TBL20] L. Touzé-Peiffer, A. Barberousse, and H. Le Treut. "The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research". *WIREs Climate Change* 11:4, 2020, e648. ISSN: 1757-7780, 1757-7799. DOI: 10.1002/wcc.648.

[Tel14] A. C. Telea. *Data visualization: principles and practice*. cRC Press, 2014.

[TR20] N. Teale and D. A. Robinson. "Patterns of Water Vapor Transport in the Eastern United States". *Journal of Hydrometeorology* 21:9, 1, 2020, pp. 2123–2138. ISSN: 1525-755X, 1525-7541. DOI: 10.1175/JHM-D-19-0267.1.

[Vie+21a] D. Vietinghoff, C. Heine, M. Bottinger, N. Maher, J. Jungclaus, and G. Scheuermann. "Visual Analysis of Spatio-Temporal Trends in Time-Dependent Ensemble Data Sets on the Example of the North Atlantic Oscillation". In: *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. 2021 IEEE 14th Pacific

Visualization Symposium (PacificVis). IEEE, Tianjin, China, 2021, pp. 71–80. ISBN: 978-1-66543-931-2. DOI: 10.1109/PacificVis52677.2021.00017.

[Vie+21b] D. Vietinghoff, C. Heine, M. Böttinger, and G. Scheuermann. "An Extension of Empirical Orthogonal Functions for the Analysis of Time-Dependent 2D Scalar Field Ensembles". In: *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. 2021 IEEE 14th Pacific Visualization Symposium (PacificVis). 2021, pp. 46–50. DOI: 10.1109/PacificVis52677.2021.00014.

[Vie24] D. Vietinghoff. "Critical Points of Uncertain Scalar Fields", 2024.

[WBR18] A. Wypych, B. Bochenek, and M. Różycki. "Atmospheric Moisture Content over Europe and the Northern Atlantic". *Atmosphere* 9:1, 11, 2018, p. 18. ISSN: 2073-4433. DOI: 10.3390/atmos9010018.

[Wei19] J. Weiss. "A Tutorial on the Proper Orthogonal Decomposition". In: *AIAA Aviation 2019 Forum*. AIAA Aviation 2019 Forum. American Institute of Aeronautics and Astronautics, Dallas, Texas, 17, 2019. ISBN: 978-1-62410-589-0. DOI: 10.2514/6.2019-3333.

[WMK13] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. "Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles". *IEEE Transactions on Visualization and Computer Graphics* 19:12, 2013, pp. 2713–2722. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.143.

[Yan+22] Y. Yang, C. Liu, N. Ou, X. Liao, N. Cao, N. Chen, L. Jin, R. Zheng, K. Yang, and Q. Su. "Moisture Transport and Contribution to the Continental Precipitation". *Atmosphere* 13:10, 16, 2022, p. 1694. ISSN: 2073-4433. DOI: 10.3390/atmos13101694.

[Yao+13] S. Yao, Q. Huang, Y. Zhang, and X. Zhou. "The simulation of water vapor transport in East Asia using a regional air–sea coupled model". *Journal of Geophysical Research: Atmospheres* 118:4, 27, 2013, pp. 1585–1600. ISSN: 2169-897X, 2169-8996. DOI: 10.1002/jgrd.50089.

[Zha+21] N. Zhao, A. Manda, X. Guo, K. Kikuchi, T. Nasuno, M. Nakano, Y. Zhang, and B. Wang. "A Lagrangian View of Moisture Transport Related to the Heavy Rainfall of July 2020 in Japan: Importance of the Moistening Over the Subtropical Regions". *Geophysical Research Letters* 48:5, 16, 2021, e2020GL091441. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2020GL091441.

[ZN98] Y. Zhu and R. E. Newell. "A Proposed Algorithm for Moisture Fluxes from Atmospheric Rivers". *Monthly Weather Review* 126:3, 1998, pp. 725–735. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2.

[Zou+18] M. Zou, S. Qiao, T. Feng, Y. Wu, and G. Feng. "The inter-decadal change in anomalous summertime water vapour transport modes over the tropical Indian Ocean–western Pacific in the mid-1980s". *International Journal of Climatology* 38:6, 2018, pp. 2672–2685. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5452.

[Zou+20] M. Zou, S. Qiao, L. Chao, D. Chen, C. Hu, Q. Li, and G. Feng. "Investigating the Interannual Variability of the Boreal Summer Water Vapor Source and Sink over the Tropical Eastern Indian Ocean-Western Pacific". *Atmosphere* 11:7, 17, 2020, p. 758. ISSN: 2073-4433. DOI: 10.3390/atmos11070758.

[ZY05] T.-J. Zhou and R.-C. Yu. "Atmospheric water vapor transport associated with typical anomalous summer rainfall patterns in China". *Journal of Geophysical Research: Atmospheres* 110, D8 27, 2005, 2004JD005413. ISSN: 0148-0227. DOI: 10.1029/2004JD005413.

List of Figures

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Ort: Datum: Unterschrift: