Universität Leipzig

Fakultät für Mathematik und Informatik
Institut für Informatik

Evolution of Moisture Transport Patterns in the North Atlantic in different Climate scenarios

Masterarbeit

Leipzig, Mai 2024

vorgelegt von

Denis Streitmatter
Studiengang Master Informatik

Betreuende Hochschullehrer:
Dr. Baldwin Nsonga
Universität Leipzig, Abteilung für Bild und Signalverarbeitung
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. The objective of my master thesis is to analyze the evolution of moisture transport patterns in various future climate scenarios. The foundation of this research lies in the MPI-GE, the Max Planck Institute Grand Ensemble Dataset, comprising an ensemble of 100 members for different RCP (climate) scenarios up until 2100. Each member provides multiple fields of relevant climate data. A challenge will be the visualization of uncertainty stemming from 100 different simulations, which will not be straightforward.

To quantify moisture transport, an integrated water vapor transport (a combination of wind and specific moisture) scalar/vector field will be generated from the MPI-GE. Windowed Empirical Orthogonal Functions (EOFs) will be used to extract spatial-temporal patterns and simplify the data, making it easier to evaluate pattern evolution over time.

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 [16, 17] humans came a long way of fighting the consequences of the increased greenhouse gas concentration in earth's atmosphere. In 1972 Sawyer summarized the knowledge and predicted quite accurately the warming at the end of the century [51]. Especially the last decades the climate crisis gained more and more attention, leading to the creation of multiple international organizations and institutions (e.g. the International Panel on Climate Change (IPCC) in 1988).

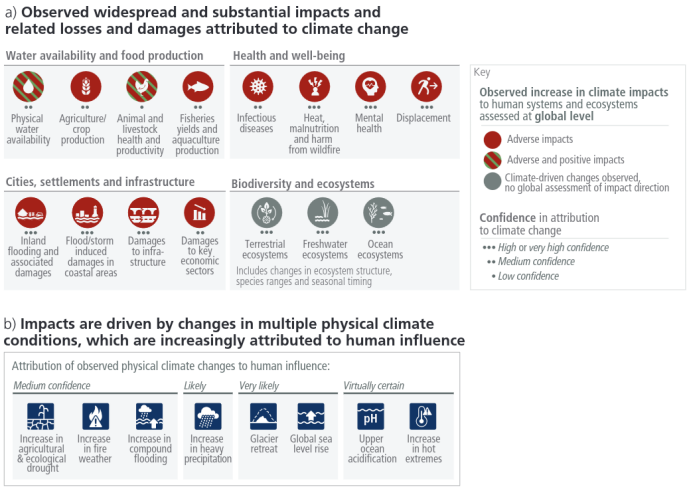

Figure 1.1: Impact of Climate Change for Humans, taken from [31]

In 2019 more than 11,000 scientists from around the world released a declaration [47], calling governments from around the world to action. The consequences for the environment and humans are prevalent and are, in part, already visible today. Figure 1.1 shows likely consequences for humans from the latest IPCC report for policymakers [31]: Flooding, malnutrition, displacement, and damages to all kinds of ecosystems can be attributed with high confidence to climate change. The sources of such consequences are manyfold,

but recent research shows that big circulation systems like the North Atlantic Oscillation[60] or the Atlantic Meridional Overturning Circulation [33] change as well.

Although the water vapor in the air accounts for only 0.001 % of the water on the earth, it is the most active part of that cycle [69]. Also, research shows that the precipitation on land does not match the evaporation, meaning the water was transported (from the oceans) to land, providing water for the ecosystems there. Analyzing the structural change of this moisture transport could help predict consequences. Motivated by the research of Vietinghoff et al., this thesis aims to evaluate similarly the systemic changes of moisture transport and precipitation patterns in Europe and the northern Atlantic.

CITE!!! Where the fuck did I read this?

1.2 Climate and Climate Research

1.2.1 Quick Overview over Climate Systems and Climate Change

In difference to weather, which is the momentary state of the atmosphere at a time, climate is the average of weather patterns over a larger period of time, usually 30 years or more [38]. 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). Earth's climate system can be seen as complex interactions of its major components: atmosphere, hydrosphere, cryosphere, lithosphere, and biosphere [26, 59]. Changes in this system can have (roughly) two reasons: Either "internal variations in form of redistributions of energy" [59], which can happen on arbitrary scales (see the discussion on the change of AMOC in [33]) or in the form of external forcings. Such forcings could be vulcanic activity, differences in solar radiation, and of course the emission of greenhouse gases (GHGs).

Figure 1.2 gives an example what effect such external forcing can have: It shows the change in effective radiative forcing (ERF) and its contributing components. ERF (in Wm^{-2}) is a way of measuring how much energy from the sun is "trapped" instead of reflected back to space (greenhouse effect). A positive value means warming, while a negative value is associated with cooling. In can be seen in Figure 1.2 that neither volcanic activity or solar radiotian changed that much, the main drivers of change in ERF are the man-made GHGs and cooling aerosols. [26]

Regarding the internal variations: Most of it is part of some oscillation, with the oscillations of the atmosphere with the hydrosphere (i.e. all liquid forms of water on earth) being responsible for large parts of climate's internal variations on decadal and interannual time frames [59]. Prominent examples for such oscillations are the El Niño Southern

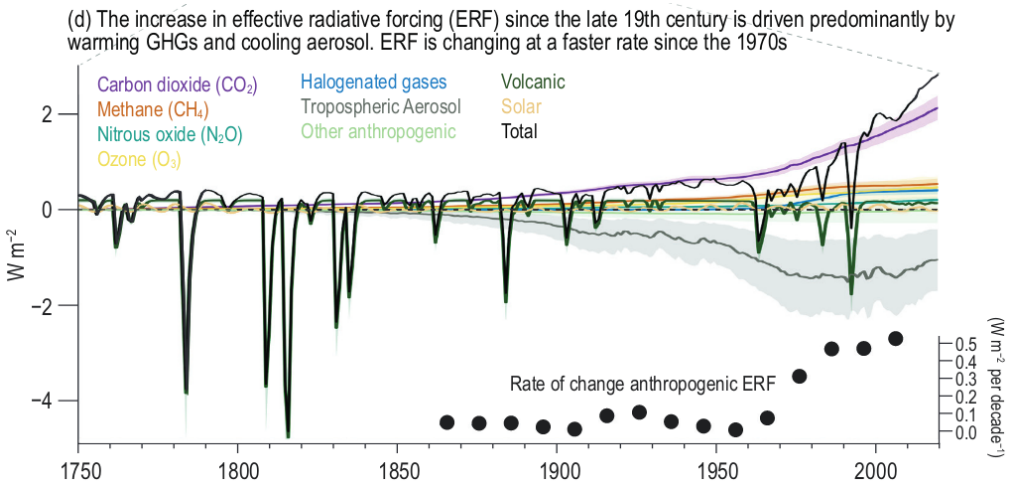

Figure 1.2: The evolution of the effective radiative forcing and contributing components, taken from [26]

Oscillation (ENSO) or the North Atlantic Oscillation (NAO), which is especially relevant for this thesis.

1.2.2 The North Atlantic Oscillation

The aforementioned NAO is "... one of the most recurrent and prominent patterns of atmospheric circulation variability" [25]. It is also one of the oldest known weather patterns, since some descriptions of Scandinavians exist from a few 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 the 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.

Explain physical modes!

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. It's basis is a characteristic dipole in the Sea Level Pressure Field (SLP) 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 the maritime air from the Atlantic towards Europe (see Figure 1.3) [25, 59]. 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 therefor to colder winters.

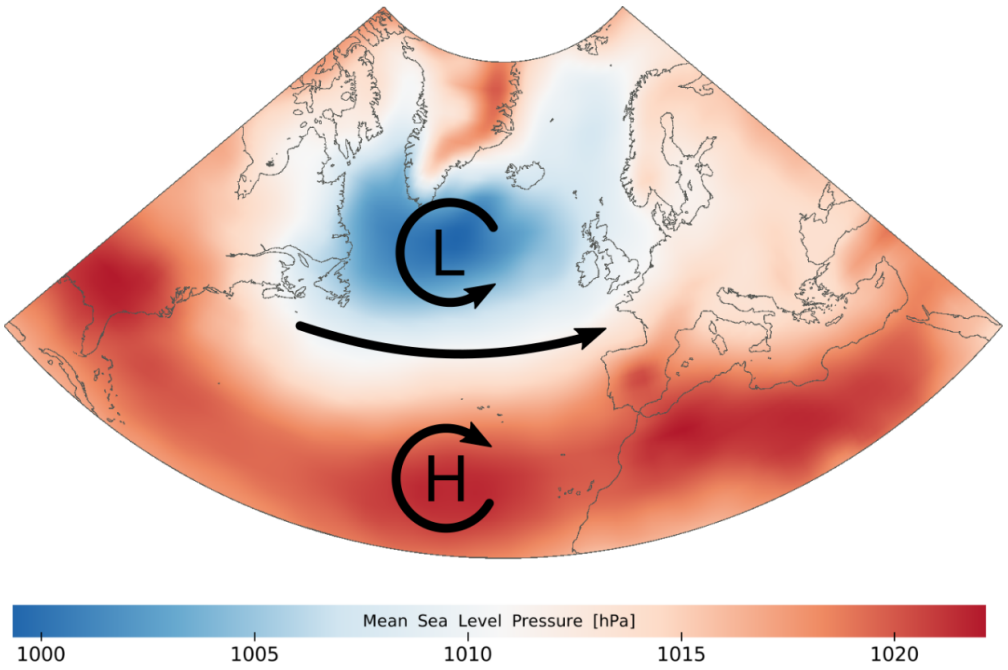

Figure 1.3: Characteristic mean SLP field of a boreal winter season, taken from [59]. It shows a high (H)/low (L) pressure pattern, directing air from the Atlantic westwards towards Europe.

This varies on an interannual scale, and this effect is called the North Atlantic Oscillation. Figure 1.4 shows the actual index based on measurements of weather stations in Iceland and the Azores (top row), which was the usual way of 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) are very similar to the measured Index (middle row).

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 to similar anomalies in the opposite direction in the later years. Therefor, Hurrell et al. points out the need to study the relationship of anthropogenic climate change and the NAO. Following this, the motivation of thesis [60] by Vietinghoff et al. tried to track the shift of the centers of the dipoles in different climate scenarios.

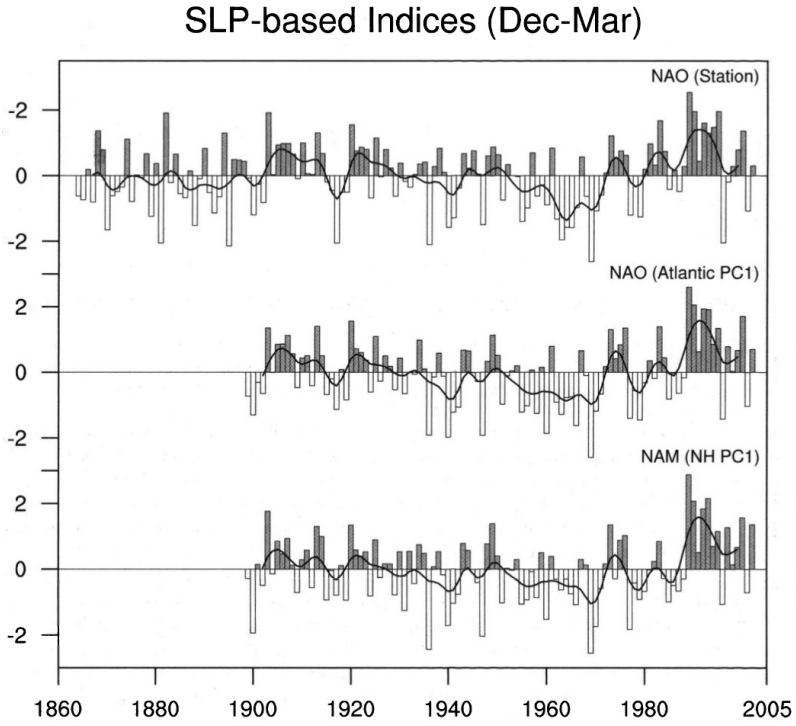

Figure 1.4: Comparison of the NAO index from [25]: The top panel are differences of SLP from weather stations in Portugal and Iceland, the middle panel is the first principal component of corresponding to the first EOF of the northern Atlantic SLP field, and the bottom panel is the same as the middle panel but for the whole Northern Hemisphere. See [25] for a more detailed description

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

The reason for the endorsement of the IPCC by the UN General Assembly 1988 was to prepare comprehensive reviews and report about the current state of scientific knowledge and research. Since then there were six assessment cycles and six reports were published, condensing the research of the scientific community. Figure 1.1 is a graphic from the latest report for policymakers from 2023 [31], displaying the probable consequences for humans in climate change.

A main source for such figures in the reports are so-called Global Coupled Models (GCMs)[1], trying to model the state and evolution of certain fields of earth data. They consist of multiple Models, each representing a major part of Earth's complex climate

[1]Unfortunately, Global Coupled Models share their acronym with General Circulation Models, which are quite similar

system (like atmosphere, hydrosphere, etc.), also allowing to model the dynamic interactions between these parts [59]. In the mid 90s the Coupled Model Intercomparison Project (CMIP) was brought to life, with the aim of streamlining results of GCMs and making them comparable. CMIP provides the outer structure, amongst others what kind of simulations to produce (e.g. preindustrial control simulations, future scenarios etc.), what kinds of fields should be generated, what kind of resolutions to provide and also how these results should be serialized. Since then the results of CMIP played an increasingly major part in the reports of the IPCC [58], and are now even called "... one of the foundational elements of climate science" [13]. CMIP is currently in its 6th phase, corresponding to the recently finished 6th Assessment Report of the IPCC [31]. The 6th phase describes an inner core (DECK + historical simulations), required for participating in CMIP, and some endorsed Model intercomparison Projects (MIPs). The latter are optional and consist of e.g. ScenarioMIP (future scenario simulations), HighResMIP (for exploring Models with higher resolutions), and GeoMIP (exploring effects of geoengineering).

The simulations are usually set up as so-called ensemble simulations, consisting of different members which are a simulation themself. The members use the same forcings, but different starting conditions and are independet from each other. Using multiple simulations makes it possible to separate the internal variability from the responses to the external forcing, enabling researchers to better quantify the consequences of climate change (for example). Additionally, it makes the research of extreme weather phenomena (e.g. droughts, floods etc.) more robust in spite of their rare occurrences [35]. This results in the challenge of working with more than one field (see Section 2.1) and visualizing the variability introduced by multiple members.

1.3 Research Questions and Thesis Structure

Following up the previous sections, the research question for this thesis is:

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

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

Can I say it like this?

M1 Generate the patterns of moisture transport and other variables

For this moisture transport needs to be somehow quantified, and this quantification needs to be calculated based on the data available in the chosen dataset (Chapter 3). Furthermore, a similar sliding window approach as in [60] needs to be implemented to study the evolution of the patterns.

M2 Study the relationships with other variables and patterns

To get a grasp of the meaning of the 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 (like the East Atlantic Pattern, 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 the 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 visualized so that they can properly be interpreted. Important is here to visualize the variability introduced by uncertain fields (introduced by multiple members of the ensemble simulation). This includes choosing a feature in the results on which this variability as well as the change over time can be analyzed. One interesteing thing to find if IVT (or other variables) experience a similar shift to the north similar to the results of Vietinghoff et al. [60].

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

2 Basics

2.1 Sampled Data, Grids and (Uncertain) Fields

The goal of this section is to give insights into the structure of scientific (sampled) data, grids, interpolation, and (uncertain) fields. Especially the first parts of this section are largely based on the work of Telea [57], for a more extensive introduction please refer to it.

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 stems from nature and is measured in physical units like kg, $\frac{m}{s}$ or similar. Continuous data conforms to the Cauchy-Criterion (also called the $\epsilon - \delta$-Criterion), which essentially 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 [57] for the mathematical definition). Continuous data can be mathematically represented as a function in the form of [57]:

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

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 C domain. Functions with $c = 1$ are called Scalar Fields, assigning 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 every position a vector in the form of $(x_1, x_2) \in \mathbb{R}^2$ or $(x_1, x_2, x_3) \in \mathbb{R}^3$, which can (but does not have to) depend on the original function domain. There are also fields related to higher dimensions (Tensor fields), but they are beyond the scope of this Thesis.

Although this kind of data is continuous in the real world, in its computational representation it is nearly always represented in a discrete way. The reason for this is that it is a) hard to achieve continuous data and b) many mathematical operations (e.g. filtering,

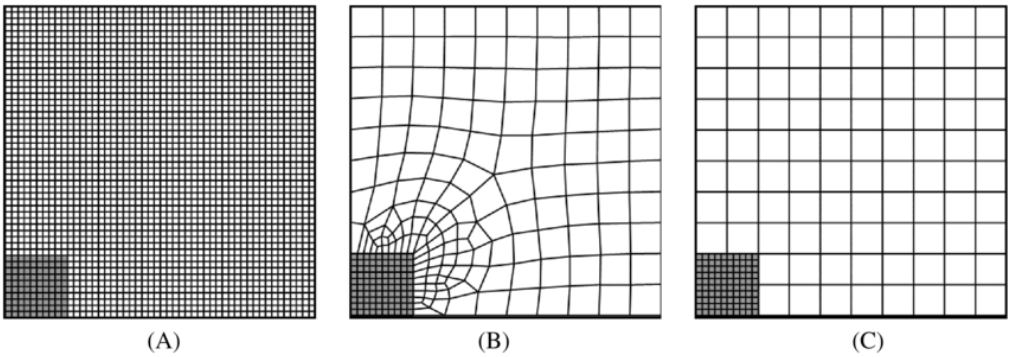

Figure 2.1: Different types of Grids: A) Uniform Grid, B) unstructured grid, and C) no-conforming grid from [28]

denoising, rendering) are hard to perform on continuous data. According to Telea [57], this 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 by using interpolation. Therefor, when a field is mentioned in this Thesis, it usually refers to such an approximation of a continuous field like in Equation 2.1.

Interpolation usually uses the structure of data, which is mostly called a *grid* (or sometimes 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 essentially an axis-aligned box spanning over 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 sample points p_i can be described with:

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

Therefor, every sample point can be described by its integer coordinates $n_1, \dots, n_d$. The number of sample points on 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 its 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 sampled points p_i and associated values f_i. In general, there are multiple ways of interpolating, for example using by using *nearest-neighbor interpolation*, assigning each point the value of the closest cell center. While this is computationally efficient, it is a staircase-like, discontinuous approximation of the original data. A better, continuous approach is linear interpolation, which interpolates the value of a point $x \in D$ based on the surrounding cell. 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 Telea [57].

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 (d), while the topological dimension refers to the actual function domain D itself, which is $s \leq d$ [57]. 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 also the earths surface is (approximately) a sphere's surface, the geometrical dimension of such datasets is $d = 3$. But since the earths surface can be is a (curved) plane, the topological dimension of such datasets is $s = 2$, which is also the reason why it is enough to access any gridpoint on earths 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.

Unfortunately, this requires the mapping of a 2D paper plane to a 2D sphere surface, which is topologically impossible (without distortions) [59]. Therefor, 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. As pointed out by Vietinghoff [59], a different map projections may fit the area of interest better, the Mercator projection in Figure 2.2 (left) is still chosen for this thesis, due to limitations in the employed map projections library (see Section 5.4). This distortion also plays a role in calculations with uniform lon/lat-grids, since coordinates in the far north (and south) are vastly overrepresented, since they actually refer to a far smaller portion of the earth's surface than data near the equator (see Section 2.2 for geographical weighting).

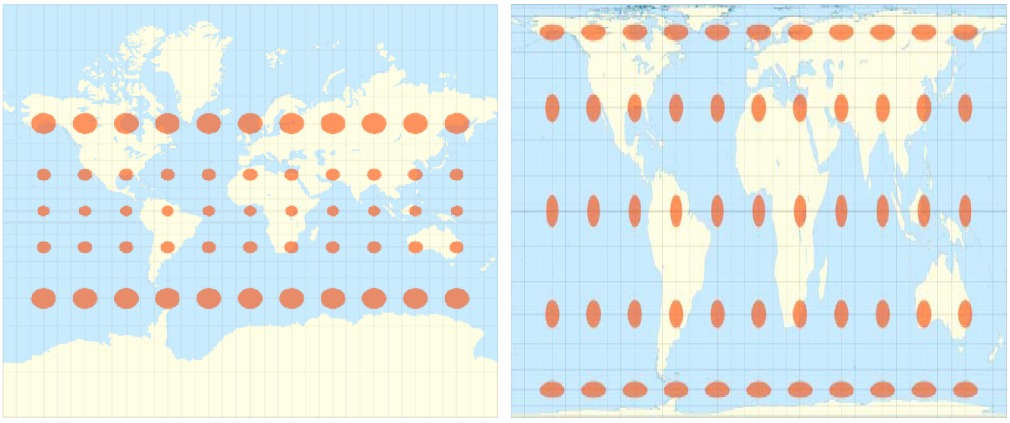

Figure 2.2: Two different map projections with different kinds of warping, depending on the map projection. The amount of warping is indicated by Tissot's indicatrices (orange circles). Left: Mercator Projection, Right: Lambert Equal-Area Projection [19]

2.1.3 Uncertain Fields

Explain uncertain fields and ensembles from [43, 59]

2.2 Empirical Orthogonal Functions

2.2.1 Overview

Empirical Orthogonal Functions (short: EOFs) analysis, also known as geographically weighted PCA or Proper Orthogonal Decomposition [59], "is among the most widely and extensively used methods in atmospheric science" [23]. One of its goals is to reduce the usually very high dimensionality of atmospheric data and can be used to link certain modes/patterns to the physics/dynamics of the analyzed system. 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 [23, 59]. The naming of the components is for from being consistent: The spatial patterns are also called spatial modes, PC loadings, EOFs or even sometimes PCs, while the temporal coefficients are also named principal components (PCs), EOF amplitudes or EOF (expansion) coefficients [23]. 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 spatial pattern [23].

This could be achieved with multiple kinds of decomposition, but EOF tries finding new sets of variables ($c_k(t)$ and $u_k(s)$ from Equation 2.4) that each capture a maximum possible amount of variance/variability of the original dataset. So the first of M modes captures the most variance, the second one the second most and so on.

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. [23] as well as their actual practical computation. For a more in depth history and derivation, please refer to [23] and their references, while Weiss [61] gives a great hands-on tutorial on POD/EOFs and their interpretation and computation.

As already explained, 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 grid point 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 usually to remove the climatology of the dataset to turn it into anomaly maps. The climatology is usually defined as the temporal mean $\bar{x}$ of the analyzed datachunk, so

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

$$\bar{x} = (\bar{x_1}, ..., \bar{x_p})^T. \tag{2.6}$$

So the values of anomaly maps x'_{ij} at each grid point are given as the departure 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 for generating EOFs is the covariance matrix defined by

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

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

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

So 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 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 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 leaves room of choosing scale and direction in a useful way (see Section 5.3 for a practical implementation) [59].

2.2.3 Calculation and Application to the geographical Domain

Since geographical data is usually given on a regular 2D grid which depicts the earth's surface, the influence of grid point density (same degree resolution is far more sparse in equatorial regions than in the Arctic) need to be corrected with geographical weights. Those can be approximated by the square root of the cosine of the respective latitude [22, 59] with a similar diagonal matrix as depicted in [22]:

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

Fortunately, there is no need to calculate the covariance matrix and solve the eigenvalue problem. Linear Algebra provides a tool called *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 (like in Vietinghoff [59]):

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

When using SVD of $\tilde{X}$ with time as the first dimension (like depicted here), 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, this result can be scaled, which is explained in detail in Section 5.3.

go over this whole section and check everything for being correct!!!!

3 MPI GE CMIP6

3.1 Overview

The dataset chosen for this project is the *Max Planck Institute Grand Ensemble CMIP6* (from now on MPI-GE CMIP6), presented by Olonscheck et al. It is a Single-model initial-condition large ensemble (in short: SMILE) consisting of multiple, coupled models: ECHAM6 for the atmosphere directly coupled to JSBACH for land and MPIOM for sea and sea-ice. The models are coupled once a day, meaning that the simulation results of the different models serve as inputs for the other models. As an ensemble simulation, it consists of multiple members, which are different variants of the simulation with the same forcings (like GHGs) but different initial conditions. For this 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. [41]

Differences to its predecessor MPI GE [36] (which was used in the work of Vietinghoff [59]) include improved time resolution (from monthly means to 3/6 hourly 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 [13]), it implements the DECK core with (amongst others) a quasi-stationary preindustrial control simulation and the 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).

3.2 ScenarioMIP: Future Scenarios and Shared Socioeconomic Pathways

Since the goal of this thesis is to evaluate the future patterns of climate change, simulations of the future are necessary. Furtunately, CMIP (Phase 3) introduced a project of future climate scenarios (ScenarioMIP) in the 2000s, which define and simulate developments of different anthropogenic drivers of climate change [40]. They play an important role in climate research and are since then the source for many figures and assessments

in IPCC reports [58]. 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" [40]. Basically, the differences between them are the forcings introduced by multiple variables, including change of land use, climate change mitigation policies, energy usage, population, economic growth and emissions [46]. For CMIP6 they extended the old model of RCPs (Representative Concentration Pathways), which were used for CMIP5, by adding so-called Shared Socioeconomic Pathways (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 across and within countries. Consumption is oriented toward low material growth and lower resource and energy intensity." [46]

while 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." [46]

These narratives are then quantified in multiple dimensions (resource availability, technical development, lifestyle changes, population, economic activity etc.). These quantifications then serve as an input for a range of integrated assessment models (IAMs), which turn them into the actual forcings needed (e.g. land and energy use, emissions) [46].

In the actual scenarios these pathways are combined with the additional radiative forcing (RCP, the earlier version of scenarios in CMIP5), resulting in a matrix which can be seen in Figure 3.1. RCP describes the level of radiative forcing (in Wm^{-2}) reached in the year 2100 (see Section 1.2). Although there are now 35 possible scenarios, O'Neill et al. defined two different tiers of scenarios ranked by their importance. Figure 3.1 lists Tier 1, which are scenarios mostly compareable to the old RCP scenarios. These scenarios are available in the MPI GE CMIP6, amongst some of Tier 2. [5, 40, 46]

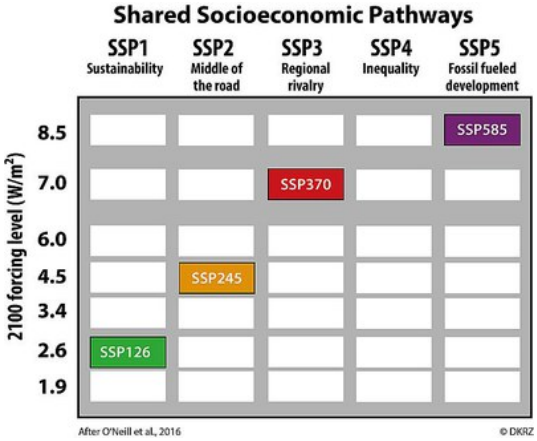

Figure 3.1: Combinations of SSPs and RCPs leading to scenarios compareable to the old RCPs [5]

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 resolution 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 for ocean. With increasing spatial resolution comes decreased availability of other variables like simulation members,

covered time period, and implemented scenarios. Although [41] 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 hourly interval. A full list of the variables can be seen in [41, Table 3], the variables necessary for this thesis are listed in Table 3.1.

Table 3.1: Variables necessary for this thesis, derived from [41]

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 the vertical levels, all variables were not available in fixed pressure layers but in so-called *hybrid sigma pressure coordinates*. In comparison to fixed pressure layers (like 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 [11]:

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

Here $p \in [P_S, P_{top}]$ is a pressure level. It was proposed that instead giving it at pure fractional levels, it would be better to smoothly converge from terrain following fractions (sigma levels) at lower (meaning near the earth surface) levels to isobaric (= same pressure) levels in higher altitudes. This gives numerical as well as practical advantages. So there were alternative functions tested for $h(p, P_S)$, but a final form[1] for calculating pressure levels at any discrete sigma level $\tilde{\eta}$ is

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

with $A(\tilde{\eta})$ being a vertical shift, which is close to 0 at low levels, and $B(\tilde{\eta})$ being a fraction of the pressure range $(P_S - P_{top})$.

[1] The equations that lead to to that can be seen in [11]

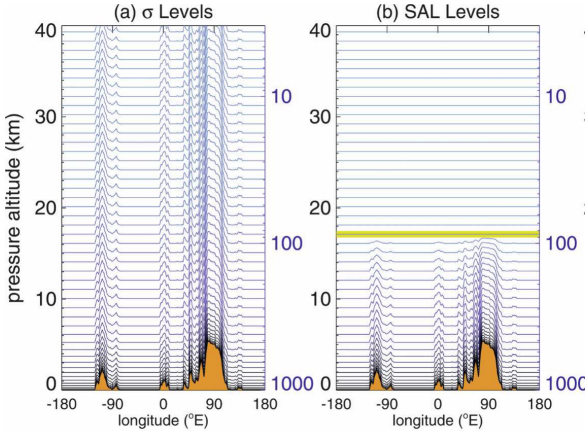

Figure 3.2: Examples of (hybrid) sigma pressure layers. a) shows sigma layers like in Equation 3.1, while b) shows a hybrid approach in the form of Equation 3.2 [11]

This results in levels of equal pressure thickness, converging to isobaric layers at higher altitudes. Which exact form $A(\tilde{\eta})$ and $B(\tilde{\eta})$ took in the MPI GE CMIP6 could not be found in [41], but every dataset using hybrid sigma pressure levels contains the variabels $ap(lev)$, $b(lev)$ and the field of pressure levels $ps(lon, lat, time)$, from which the pressure at any grid-point and level can be calculated with

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

The top pressure can be ignored in this dataset since the upper border is zero.

3.3.3 Structure of the data

The data is available in the HPC cluster of the DKRZ[2], the structure can be seen in if Figure 3.3.

It starts in a root folder for one MIP, being ScenarioMIP or the CMIP core, followed by the institution and the 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 in turn is based on HDF5 [15]) and split up in timescopes of up to 20 years. The version is named after the date of the simulation, and since a later version indicates a problem fix in the older version, it is generally advisable to pick the latest version for each variable.

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

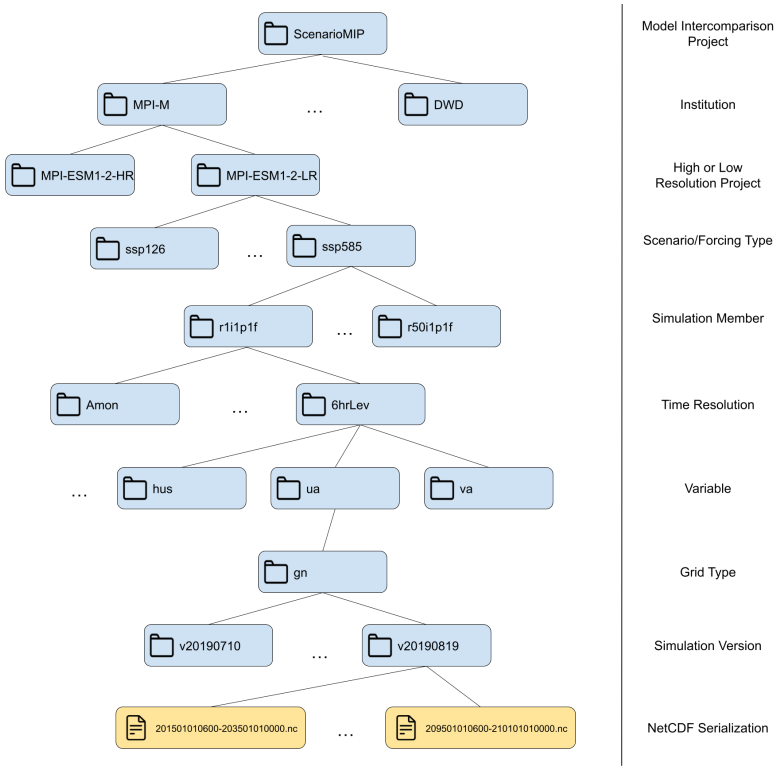

Figure 3.3: Structure the data is available on the DKRZ cluster. Example is given for ScenarioMIP, but applies aswell for CMIP DECK for historical and piControl.

3.3.4 NetCDF Datasets

The goal of the Network Common Data Format (NetCDF) was to create a machine indepedent 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 (like physical unit, other names, important notes). [45]

Figure 3.4 gives an example of said structure in 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

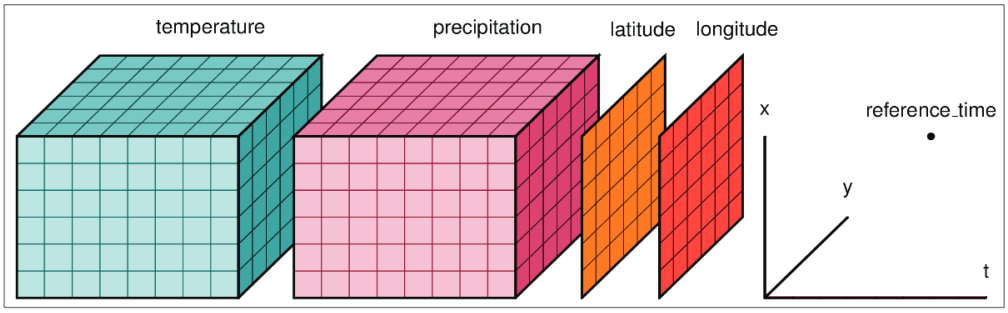

Figure 3.4: An example of a named multidimensional dataset from [24]: Precipitation and temperature are the variables while x, y and t are dimensions. Longitude and latitude are coordinates (also variables) defined by x and y.

of the grid, giving a reference for the location of the grid. The dimensions of the dataset are x, y and t, giving the both variables the shape (x, y, t). [24]

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 [60] and the affiliated dissertation [59], which tackles the issue of detecting critical points in unstable scalar fields. Hereby [60] analyzes the MPI GE [36] from the 5th phase of CMIP, an ensemble simulation with 50 members. The goal was to find the probable centers of pressure high/lows in the NAO pattern (see Section 1.2) and to track their shift over time. They employed a sliding window approach, computing the dominant pattern (see Section 2.2) for each window and member, and determine 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 the pressure high and low. The results show that the patterns do change, and this change is more pronounced if climate change is stronger. Also, there is no significant change if the climate remains stable.

4.2 Moisture Transport

To computationally extract any spatio-temporal patterns of moisture (transport), it first needs to be quantified in any way. The variable from the MPI GE CMIP6 used for this task is the *specific humidity*, which has no unit and is a float value between 0.0 and 1.0, denoting the percentage of water in the air at a specific grid point. The vast majority of literature regarding moisture transport use some form of vertically integrated humidity and the variants will be explained in the following section. A popular usage of these

quantifications was to find a filamentary weather structure called "Atmospheric Rivers"[1], a prominent way of water vapor transportation in the extratropic regions [20].

The most straightforward way of quantifying moisture is **Vertically Integrated Water Vapor (IWV)** [2, 20, 37, 52, 66], which is essentially the vertical integral of the specific humidity q over the pressure levels p from earth's surface P_S to some upper limit in the atmosphere:

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

Similar to Equation 4.1, Zhu and Newell proposed in [68] to use **Vertical Integrated Moisture Transport (IVT)** for Atmospheric River detection. It is calculated by vertically integrating over the different pressure levels the zonal (along latitude lines) and meridional (along longitude lines) fluxes. It became a popular metric for finding atmospheric rivers [20], sometimes alongside IWV [12]. 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 [14]. Here u and v stand for the zonal and meridional components of the horizontal wind vector. An equivalent way is to calculate the zonal and meridional component 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 said vector field

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

is also a popular choice in detecting atmospheric rivers [44, 55] and other use cases [1].

The IVT is also part of the atmospheric moisture budget [64] (and similar in [54]) given by

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

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

With E being the total evaporation and P the precipitation. Yang et al. showed in their report [64] the directions of moisture flux and its evolution in the last three decades. The analysis was done for all continental borders based on the big ERA5 reanalysis. The metrics used for this analysis were mostly the evaporation E, precipitation P and the moisture transport convergence $VIMC = \frac{1}{g}\int_0^{P_s} \nabla \cdot q \begin{pmatrix} u \\ v \end{pmatrix} \mathrm{d}\ p$ from Equation 4.6.

While the integration in the previous equations integrates from the surface to the outer border of the atmosphere (0 Pa), it is quite common to integrate up until the limit of 300 hPa [1, 21, 30, 68], since the amount of moisture in the regions from 300 hPa to 0 Pa is quite neglible and amounts in total to about 2-3 cm/year in terms of freshwater flux [67].

There are also some other notable other algorithms, namely stable oxygen isotope investigation [34] and langragian backwards trajectories [66], but both rather look for the origin of the water vapor instead of its destination and are therefor out of scope for this thesis.

4.3 Pattern analysis regarding IVT

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

1. Generate the EOF patterns alongside their temporal patterns. Usually those are visualized for an overview.
2. Use other variables 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 [1, 30] or the raw data of precipitation.
3. Visualize the results using maps.

An overview of datasets, timescopes and other metadata is given in Table 4.1.

Published in 1982, Salstein et al. provided the first example of calculating EOF on IVT. Based on data from 91 weather stations, they computed the IVT of the whole Northern Hemisphere. Statistical significance was determined by employing a Monte Carlo testing method. The EOFs were computed on the IWV, the zonal and meridional IVT fields respectively, but they also evaluated an approach of combining both IVT components in one

data vector. They reported the significance of the primary mode of IWV, encoding nearly half (44 %) of variance of the data.

Although most found related work uses EOF analysis, Teale and Robinson employ an approach using Self Organizing Maps (SOMs) to detect patterns of moisture transport in the eastern United States. SOMs are a machine learning approach to reduce data dimensionality, producing a 2D map of higher dimensional data. While they acknowledge the efficiency of EOF to extract 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. But despite the higher moisture content, fluxes with lower moisture transport dominate water vapor movement due to their prevalence. Many of these fluxes meet typical criteria for atmospheric rivers, with varying trajectories and sources suggesting diverse 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 [56].

CITE?

Ayantobo et al. analyzed the primary six modes of EOF in China, which was grouped in 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. It was shown by comparing the temporal patterns of the primary mode of EOF with the ENSO, that these patterns were related. The cross-wavelet coherence revealed that IVT and ENSO time-series were coherent, which implies that increased IVT was prevalent linked to increased ENSO activities [1].

Wypych et al. compares the patterns of perceptible water (IWV) in Europe for different seasons/months for the last ≈ 50 years. Similar to [1], Europe was grouped into different regions with different moisture conditions. This revealed significantly different moisture patterns for the regions, for example Northern Continental vs. Northern Atlantic. The results confirmed the expected important role of atmospheric circulation for the moisture in the winter by measuring the correlation, while relationships were substantially weaker for transitional months like April or October. Noticable was also the lack of correlation between the atmospheric circulation and the moisture patterns in summer.

Fernández et al. analyzed the precipitation modes in the Mediterranean Sea and linking them to the moisture transport in the same area. A goal of this analysis was to contribute to the understanding of the reduction of precipitation which happened in the area as well as to the low-frequency precipitation variability, leading to multiyear drought periods. They employed multiple methods of validating their data: The precipitation data as well as 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, comparing the variability of actual data with random test data, while degeneracy of the EOF modes was tested using the method of North et al. [39]. Results of the analysis identify the interpretation of the three main precipitation modes: The first mode (22 % variance) seems to be linked to the NAO and 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 increased precipitation in the northern part of the domain. Additionally, moisture transport during positive and negative phases of leading mode showed increased inflow of moisture from the west [14].

Similar to [14], Zhou and Yu analyzed the anomalous summer rainfall patterns over China and link them to water vapor transport. They confirmed their results by using a second dataset for IVT calculation. They 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 to moisture transport, they identified the different ways how these heavier rain areas are coming about by certain convergences of water vapor transports. Furthermore, they compared the supply of anomalous rainfall patterns to the one of normal monsoon rainfall, revealing that those differ significantly [67].

In [21], the authors calculate rotated EOF on IVT data and try to analyze the relation between the 15 most dominant modes and the occurrence of atmospheric rivers (AR) on the USA west coast. 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 a few modes seem very influential for certain regions' AR activity, while others seem to play no role at all. They also identified favorable and unfavorable circulation states (e.g. amongst others a low pressure anomaly in a certain region) for AR occurrence [21].

Kim and Alexander showed in their analysis the connection of the IVT patterns in the western USA to three different ENSO events (eastern pacific El Niño (EPEN), central pacific El Niño (CPEN) and La Niña (NINA)). While EPEN events are associated with large positive IVT anomalies from the subtropical Pacific to the north-western USA, CPEN events lead to enhanced moisture transport to the southern USA. During NINA events the mean IVT anomaly is flipped in comparison to EPEN and CPEN. Furthermore it was shown that IVT patterns computed for these events differ significantly from the ones computed for neutral years. Furthermore the results were connected to precipitation anomalies on the USA west coast, showing huge differences (especially for the northern part of the coast) for EPEN

and CPEN events. But the authors also emphasize that while the suggestions are strong, exceptions occur (e.g. one El Niño leading to a dry winter, another to the opposite) and need to be studied in greater detail.

Similar to [60] 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 relation to sea surface temperature (SST) revealed that the correlation between the mode and SST rose in the mid 80s, from weakly correlated to significant positive correlation between IVT and SST anomalies. Furthermore, the primary mode seems to be regulated significantly bei ENSO. The second most significant mode is related to the variability of the tropical Indian Ocean dipole (defined by the differences in average SST) [70].

A different approach was employed by [69], evaluating the EOF patterns of vertically integrated apparent moisture sink. 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 the ENSO in the previous winter season, while the second mode seems to originates from internal atmopheric variability. Based on the much higher standard deviations in ENSO years, it seems that 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.

While the main focus of [65] 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 from the regional coupled model aligns better with results from other datasets and reality than the regional uncoupled model.

Li and Zhou evaluated the connection of the IVT-EOF patterns to ENSO in the asian western northern Pacific. They used a different approach then 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 a 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 [32].

Check again, fill gaps and improve Variable used for eof (maybe also add here compared vars, like SST and so on)

Table 4.1: Overview table of patterns with moisture transport

Release Year	Pattern extraction	Area of Interest	Timescope	Time Resolution	Studied Season	Variable used for EOF
2020 [56]	SOMs	USA east	1979 to 2017	daily	all year	IVT norm
2022 [1]	EOF	China	1979 to 2010	daily	all year	IVT norm
1982 [49]	EOF	Northern hemishpere	1958 to 1973	monthly/yearly	all year	IWV, IVT_u IVT_v, combined
2003 [14]	EOF	mediterranian sea	1948 to 1996	6hr	DJF	P
2005 [67]	EOF	China	1951 to 1999	monthly	JJA	P
2018 [21]	EOF	USA (west coast)	1948 to 2017	daily	NDJF	IVT norm (assumed)
2015 [30]	EOF	western USA	1979 to 2010	6hr	DJF	IVT norm (assumed)
2018 [70]	EOF	TEIOWP	1961 to 2015	monthly	JJA	IVT
2020 [69]	EOF	TEIOWP	1958 to 2018	6hr/monthly	JJA	Integrated Water Vapor Sink
2013 [65]	EOF	East Asia	1997 to 2002		JJA	IVT_u IVT_v
2012 [32]	EOF	East Asia	1979 to 2009	monthly	summer	IVT
2018 [63]	EOF	Europe	1981 to 2015	daily	all year	IWV, SLP

4.4 Uncertainty Visualization

Since the used dataset (see 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 stemming from them. This section summarizes advances fitting for this topic, giving a frame of references of current possibilities of visualizing uncertainty.

Kamal et al. 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. They grouped all kinds of representing uncertainty in two categories: quantification, consisting of mostly mathematical approaches of handling uncertain data, and visualization, displaying the uncertain data in a way directly. An overview of the different kinds of uncertainty visualization were given: Manipulation of attributes (like shading), animation, visual variables (like color, hue, brightness), graphical techniques like box/scatter plots and glyphs. 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 [29].

A way of using animation to display uncertainty in scalar fields was shown by Coninx et al. 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 [8], evaluating effective parameters for proper presentation of the uncertain area [8].

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. Additionally, they added tools for exploring the uncertainty of datasets, like a color map of the whole dataset uncertainty. 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 ribbon width to the uncertainty of isocontours at a specific grid point. The resulting tool was tested by two meteorologists, and classified the results as beneficial [50].

Another way of visualizing groups of isocontours are contour boxplots proposed in [62], grouping isocontours together 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). But the implementation is not as straight forward 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 a isocontour to a whole set of isocontours). While the results look very promising, it lacks a publicly available implementation, making it hard to use the approach.

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. While 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. [63]). Additionally, there has been no EOF IVT analysis for ensemble-scale data, just for reanalysis data (see Table 4.1) or actual weather station data. To analyze the evolution of spatio-temporal patterns is also quite underrepresented, since most approaches apply the pattern analysis on the whole available dataset (in most cases around 50 years), exceptions are the motivational work for this thesis from Vietinghoff et al. [60] and Zou et al. [70]. Additionally, to my 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 presented approaches could be quite useful in this thesis: The animated Perlin noise from [8] 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 for the decision of the feature extraction). Unfortunately, none of these algorithms is available as a library , which hinders application to a great extent.

ref to feature selection section

double check if this is true

So this Thesis tries filling 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 [59, 70]). The variables chosen for this are of course the IVT, but also Sea Level Pressure, representing the most influential oscillations of that area (NAO and EAP), and precipitation, since it is very significant for ecological as well as 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 [65, 69, 70]. While it is easy to compute and visualize the comparison of EOF patterns of different variables, Dommenget and Latif [10] provide a good example why EOF patterns are hard to interpret and link to actual physical modes. They recommend using multiple evaluations and statistical tools (like regression) to link the mathematical modes yielded by EOF analysis with the actual existing, physical modes in the real world. While 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 (to the authors best knowledge) exists for patterns of IVT and precipitation in Europe. So to interpret the results, it is important to make sense of them by checking their relationships 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

5.1 Overview

Explain what I want to do using the CMIP6 simulations: Describe what the general plan is: Visualization of the moisture transport in Europe with the help. Also define what the goals of the visualizations are: Visualize different scenarios for comparison, visualize uncertainties of different members, visualize evolution over time, also try combining those. Here should be a graphic that explains the workflow that transforms a simulation into some nice pictures

5.2 Preprocessing

The purpose of this step is to prepare the datasets for generating the patterns via EOF and generating visualizations. For this the datasets need to be reduced to the area of interest (northern Atlantic and Europe), the directory structure shown in Figure 3.3 needs to be simplified, the necessary moisture transport (see Section 4.2) needs to be calculated and data needs to be reduced to different time resolutions (daily, monthly).

Still need to write somewhere why

The calculations were performed on the high performance computing cluster[1] of the German Climate Calculations Center (DKRZ), due to the MPI GE CMIP6 is saved there and downloading the data would take a lot of time. This also result in the goal of this step to minimize the hours on the HPC system since they get billed by the time using nodes. Although these steps seem easy, due to the large sizes of the datasets and other issues many challenges were met. In the following those will be explained with regard to the step they occurred in.

5.2.1 Chosen Framework

The goal of this step is to prepare the data for further usage. After a few failed attempts with other languages/tools (CDO and Julia, see Section 5.2.3), the Python libraries xarray [24] and dask [48] were chosen as the fitting tools for this step. Xarray is a library

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

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 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 important, 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. [48]

5.2.2 Process

The following the process for handling one timescope of member of one scenario is described. For one member the different timescope files are handled iteratively. Scaling it up for all the members is trivial by either 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.

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 is 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 are here 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 dimension 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.

The latter choice is far more important: In NetCDF datasets data is often grouped into chunks of a certain, useful size and these chunks can then be compressed to reduce disk

[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.

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 result 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 which visualizes the process of the dask task graph execution (see Figure 5.1). Grey areas in the *Bytes stored per worker* section in Figure 5.1 show memory spilled to disk, which is an indicator for too large chunk sizes. Red sections in the *Task Stream* section refer to worker-to-worker communication, which may be an indicator for 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, since too much is spilled to disk. [6]

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 argumentation, the only useful way of changing the chunks are different amounts of time snapshots per chunk. Using the dask dashboard to evaluate different chunk sizes, the optimal chunk

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

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

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 [60] and [25], 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 can't be loaded as one slice. Therefor, the longitude coordinates are first transformed to 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 either be used for further calculations (Step 3) or directly saved as a NetCDF file (Step 4).

The size of data could be further reduced in this timestep by selecting only the relevant winter months (see Section 5.3), but with future work in mind the whole year was kept in this stage.

3. Calculating IVT Field

The first step to calculate the IVT field is to convert the hybrid sigma pressure levels (see Section 3.3.2) to actual pressure values. For this Equation 3.3 is used for calculating a new variable `plev` containing the pressure values at each grid point 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. Similar to related literature [1], a constant value for the 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 to 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. Generating daily/monthly means

Since the related literature does not entirely agree regarding the timely resolution of IVT in EOFs (see Table 4.1), the six hourly data is reduced to monthly and daily means using CDO. Monthly IVT is also called stationary moisture transport and dominates the total water vapor transport [67], while the anomalies (departures from monthly mean per daily/subdaily timestep) are transient components. Both are important parts of total moisture transport, but since a comparison of stationary and transient components is beyond

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

the scope of this thesis, only monthly data will be used for the further analysis. The higher resolutions are still saved and can be used in future work.

5.2.3 Challenges of Preprocessing

The steps described in the section before were just the final attempt. The first idea was using Climate Data Operators [53], a command line tool containing multiple operators for processing 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 [18], using just a NetCDF library [3] 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 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 moved data in memory was minimized by preallocating the needed RAM and writing directly to the preallocated space. Furthermore, other NetCDF libraries were tested, but simple loading times were very similar. Although this significantly reduced

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

the amount of allocations, the effect on loading time was negligible. To actually archive a significant boost in loading time it was tried to load the required datasets (located in different files) in parallel. Unfortunately, the used library [3] encountered a segmentation fault used in multiple threads, so the alternative libraries NetCDF.jl and HDF5.jl were explored, since the HDF5 standard allows parrallel access to files [15]. Although the parallel access to files using multiple threads (with HDF5.jl) lead to increased speeds in 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 up into different processes, each one loading data from one variable. This actually reduced time spent to one third in tests, but testing it on the 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.

5.3 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 [4]. Although it was tried using it for the preprocessing and failed (see Section 5.2.3), the reduction of the data was enough to make it easy working with the NetCDF library [3] implemented in Julia. Julia has the advantage of making it particularly easy and intuitive working with matrices and high dimensional arrays. Additionally, it is (usually) also very computationally efficient doing so, which is the reason it was chosen for the implementation of this step. Other factors are the great reproducibility of code and the fact the Makie framework (based on Julia) was chosen for the visual analysis step (see Section 5.4), which made it possible to reuse some already written code.

In general, the procedure of this step is based on the approach of [60], 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 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 [60], keeping only the months December, January, February and March (in the following DJFM). So during loading the irrelevant months are filtered out, additionally 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 with the 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. Then the data is grouped in certain scopes for the sliding window approach. For this 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. Following the argumentation of Vietinghoff et al. [60], the window size affects the smoothing of the data: The larger the time window, the better the noise is smoothed out. But 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. [60]. Similar to Vietinghoff et al., scopes of 30 and 50 winters are implemented. The next scope is then shifted by one winter season, and again counted for 30/50 seasons. In difference to [60], 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. These steps are then computed the same for any other time scope of any member of any variable.

1. Preparation of the Data for SVD

Before calculating SVD, the data needs to be prepared accordingly. After the scoping, the data for each scope is available as a Following the argumentation from [59] 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,

[9] m being the time dimension size

so the time is the first dimension. Now the shape of the data is $(time, spat)$, and SVD can be calculated.

2. Calculating EOFs with SVD

In the next step the Singular Value Decomposition is applied to the geographically weighted two-dimensional data chunk. For this the SVD implementation of Julia's `LinearAlgebra` package (which is part of the Standard Library) is used. This approach computes all m[9] modes, although at most the top five modes are used in this thesis. This was feasible for the monthly data used in this thesis, but once this approach is scaled up to daily or subdaily data it should be replaced by faster algorithms like Snapshot POD or the SLEPC implementation used by Vietinghoff [59]. 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 compute the eigenvalues (σ_i^2), which are the share of variability encoded in that corresponding mode i with

$$V_i = \frac{\sigma_i^2}{\sum_{i=1}^{m} \sigma_i^2} \tag{5.2}$$

Then the modes are cut of at a limit n (usually five) to reduce disk space usage and loading times in the visualization processes later.

3. Post Processing of the EOFs

After the first n modes of interest were cut off, a few more steps are required to be able 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 to some vector (see Section 5.3.2) and the weighting is reverted 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 [27]. For this the five most significant modes of the spatial and temporal patterns are persisted, alongside the singular values (for scaling and variability computation) and the sum of all eigenvalues 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 the 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 5.3.1, the latter in Section 5.3.2.

5.3.1 EOF Mode Scaling

As explained by Vietinghoff [59], there are two particularly useful ways of scaling the results: Either by multiplying the temporal patterns (or 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 EOFs in the original unit of measurement. Depending on the type of used visualization, the former or latter are more fitting, so in contrast to [59], the scaling is done at the loading of the data to be more flexible.

5.3.2 EOF Mode Alignment

Since basically only two scaling modes are useful (see Section 5.3.1), the sign/direction of that vector is a last, but very important choice to make. While 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 maximum of a few modes (since most don't compute multiple modes, and none use a 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 TODO patterns to align for one scenario, it's 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 [59]. Its solution for this is providing a field F to which we compare our spatial mode and decide to whether flip it (≡ multiplying it with -1) or not.

The method chosen by Vietinghoff [59] is to use the Scalar/Dot 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, meaning the spatial pattern is closer to the mean of the data. This process is illustrated in the Figure TODO 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 [59] (see Figure TODO).

Make fig with the mean fields, cleaned by spatial mean and the first EOF pattern

Unfortunately, this works only well for the primary EOF mode, while the rest compared to yields not so successful results. A way of testing the alignment of patterns is by comput-

ing the cross-correlation boxplots from Section 5.4.2 with and without absolute correlation values. Figure **??** gives an example of such an analysis. 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 5.2.

5.4 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 takes reference to techniques used in related work.

For this task Julia's Makie Framework [9] was chosen. It allows to easily create a complex layout for figures and allows creating 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" [9]. Additionally, there is a library[10] available for projecting data onto different map projections.

While the previous two Sections had the goal of generating the 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 the IVT EOF patterns as well as displaying the variability introduced by the multiple members of the chosen dataset (**M3**).

5.4.1 Spatial Pattern Analysis

The main focus of this thesis lies in the visualization of the spatial patterns, since those are easy to understand visually. The main challenges are here 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.

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

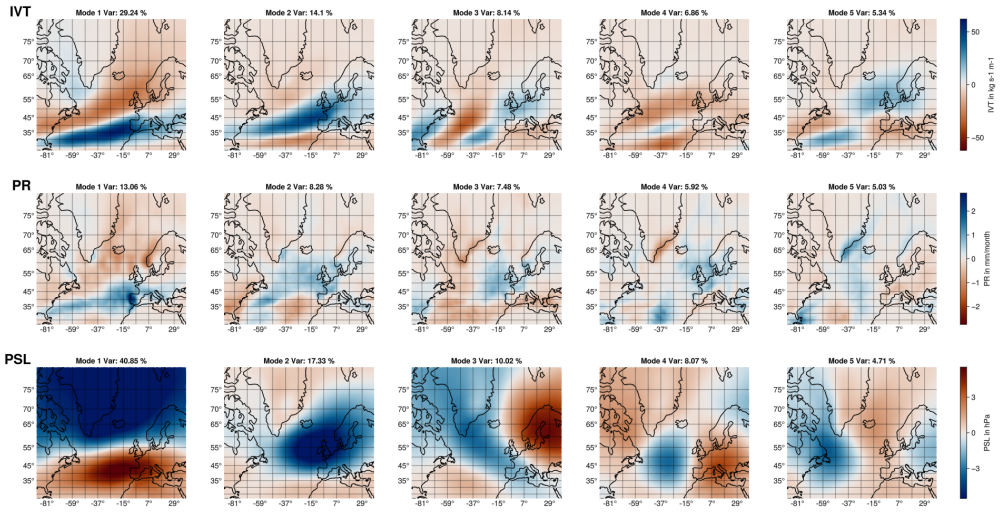

Figure 5.2: The five most significant modes of each of the processed variables (using the short names given in Table 3.1). Taken from one member with 50 winter timescope and the scope from 1868 to 1918

Although there are some examples of visualizing uncertainty in scalar fields (or features thereof like iso contours), none of them provide an existing implementation, which greatly hinders the usability. Since a new implementation of those concepts was outside the scope of this thesis, a new way of visualizing the ambiguity of multiple members needed to be found. While ways of representing uncertain scalar fields exist (like [8]), it may not always be helpful to visualize the full ambiguity of the dataset since it could generate too much visual clutter. Instead, a certain feature could be extracted to show the ensembles' variability.

Feature Selection

As shown in Section 4.4, there exist multiple ways of visualizing scalar fields, either by transforming the actual scalar field somehow (like the approach of Coninx et al. [8]) or by selecting an interesting feature in it and visualizing it. The latter approach was used by Vietinghoff [59] (using the extreme points of the fields) and by [50, 62] 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. Therefor, 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 defining the field. The best

known contour lines are lines of same height in maps of mountains. And 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. [42], but for isolines and not for isosurfaces.

Visualizing the Feature

The most straight forward way of visualizing uncertain isolines are spaghetti plots as used in the work of Sanyal et al. [50], by simply drawing all the isolines of the different members, usually in different colors. This was also implemented in this Thesis, but spaghetti plots have a lot of usual drawbacks. While they work very well for areas where the contour lines are very close to each other (see Figure **??**), they tend to overcrawded areas making the visualization overwhelming and chaotic. Additionally, contour lines give a false sense of precision in data that is actually not that precise.

Level Crossing Probabilities using Hexbins

To fix those issues of spaghetti plots, another method was implemented using hexbins. Hexbin plots[11] as used by Carr et al. [7] for geographic data are an alternative to heatmaps, depicting 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 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 distribution better than square bins (≡ heatmaps) as it was depicted in [7], as well that it can be more visually appealing to humans [7]. The main idea for this thesis was not only the better representation, but also that the number of hexagons (or their size) can be chosen freely, so by choosing a resolution similar to the underlying grid gives a sense of data precision.

The approach using hexbins was conducted as follows: First, the required contour lines were computed, which are represented as a list of 2D points. The threshold for displayed hexbins was set to one, so that areas without any contour line stay free of hexagons. Since the resolution is too high at some places (multiple observations per bin) and too low at others (no observation hitting the bin) like it is depicted in Figure 5.3 A and D, a sampling algorithm was used to sample along the line. Of course this results in a distorted color bar since one observation, but by choosing a small enough integration distance (i.e. 0.01),

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

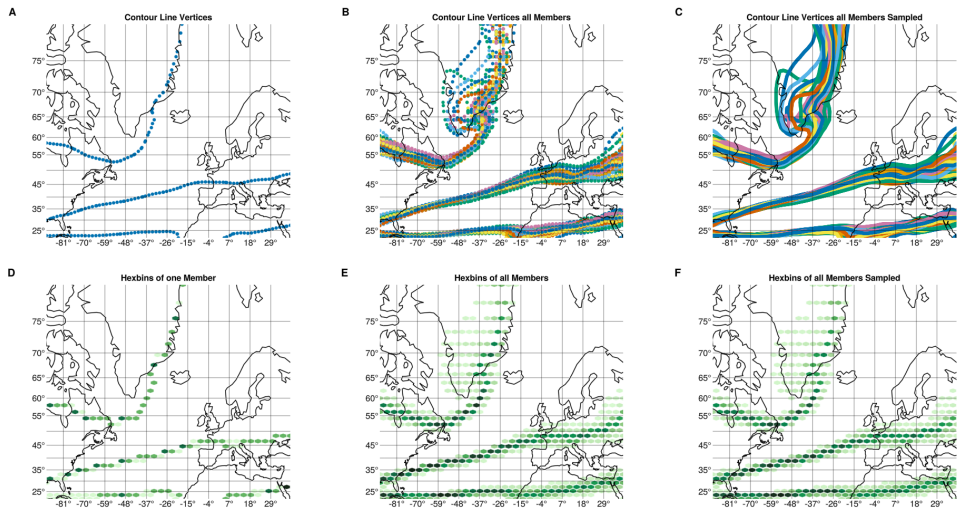

Figure 5.3: Example of the hexbin approach using the dominant IVT pattern. Sample distance is 0.1. A to C show the contour lines vertices in one member, all members and all members with sampling along the line. D to F show the hexbin visualization of the corresponding vertices.

this distortion is reduced to an invisible minimum. So now by using multiple members (Figure 5.3, E and F), it approximates the number of lines hitting a certain bin. An analysis comparing this approach to the traditional spaghetti plots is given in Section 6.1.

5.4.2 Temporal Pattern Correlation

The spatial pattern is of course only the 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 mess when using mean functions or boxplots.

So instead the temporal activity can be used to evaluate the relationship of certain patterns with other variables or even other patterns. Obviously, 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 therefor be justified separately from correlation. Cross-correlation measures the correlation between two signals with a pre-defined 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 to 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 introducing a lag to the signal, it could reveal certain temporal relationships, e.g. a positive IVT EOF coefficient of EOF2 (see Figure 5.2) leads to positive activity of the precipitation pattern positive in Great Britain (e.g. PR EOF2).

This evaluation was conducted in the following way: 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 the half of the scopes' length. Of course a greater extend of lag could be chosen, but it seemed very unlikely that moisture transport in a certain month would affect precipitation a many decades later (same reasoning for the other variables). Then the maximal extend of correlation per member was used for a boxplot of that scope. The lag associated with that value per member was then displazed in a separate boxplot of lags. Then this procedure was repeated for every scope from the begin of the historical simulation to the end of each scenario, depicting the evolution over time.

Since only the temporal coefficients (without the spatial patterns) are evaluated here, it is crucial that the alignment of the patterns (as described in Section 5.3.2) works well since it can distort the correlation results significantly. In fact, using this kind of analysis revealed the lack of stability of the procedure used by Vietinghoff et al. [60] in EOF2 and lesser modes. For this the plots of the maximal absolute value of correlation were compared to the normal maximal extend 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.

Example Picture?

Comparing an EOF pattern with another variable

Since it is notoriously hard to connect (mathematical) EOF modes to real physical modes [10, 23], 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 [10]). Instead, the recommendations of Dommenget and Latif [10] were followed,

using different statistical tools to evaluate the modes. Some of the related work [32, 67, 70] used regression maps, depicting the regression slopes between the EOF coefficient $c_a^X(t)$ of variable X and mode a (e.g. the dominant EOF mode of IVT $c_1^{IVT}(t)$)and the temporal evolution of each gridpoints $Y(lon, lat)(t)$ of another variable Y. Fernández et al. [14] used a similar procedure but with the correlation of $c_a^X(t)$ and $X(lon, lat)(t)$, so depicting the correlation coefficient for each gridpoint with the temporal pattern for the same variable X. This tackles 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." [10]

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. Therefor, they display areas with correlation higher (or lower) then an interactively chosen correlation level (e.g. 0.7). Also, the mode and the scope can be interactively chosen, to explore different modes and their evolution in time.

6 Results

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

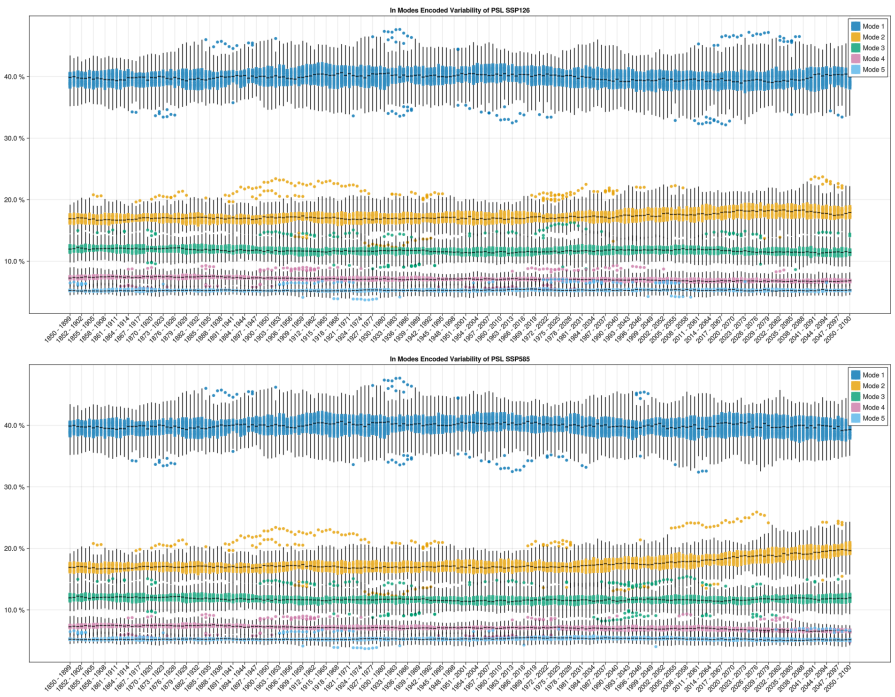

Figure 6.1: Boxplot of the variability encoded in the top five modes of PSL EOF.

The first simple evaluation is to look at the change of 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 significant 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 members of the simulations being around ±2%, with no significant trend over the years. The secondary mode (EAP) median 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. Comparing it to the SSP585 scenario, it is obvious that there is very little to no change in Modes 3-5 and 1. But interestingly, the median variability encoded by the secondary mode rises from the 17% in the 1850 - 1900 scope to around 20% in the last one, exposing a clear trend over the course of climate change.

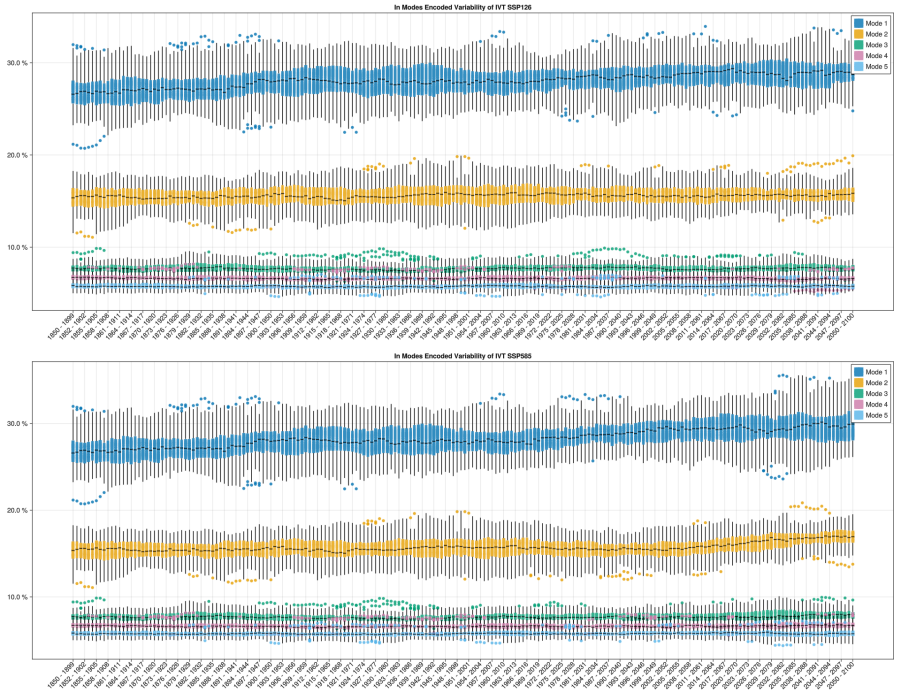

Figure 6.2: Same as Figure 6.1 but with IVT

The same analysis with the IVT patterns (Figure 6.2) reveal a general upwards 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%. Similar to Figure 6.1, the secondary mode (representing around 15% of variability) shows upward trend in scenario SSP585 to around 17%, which is not recognizable in the SSP126 scenario.

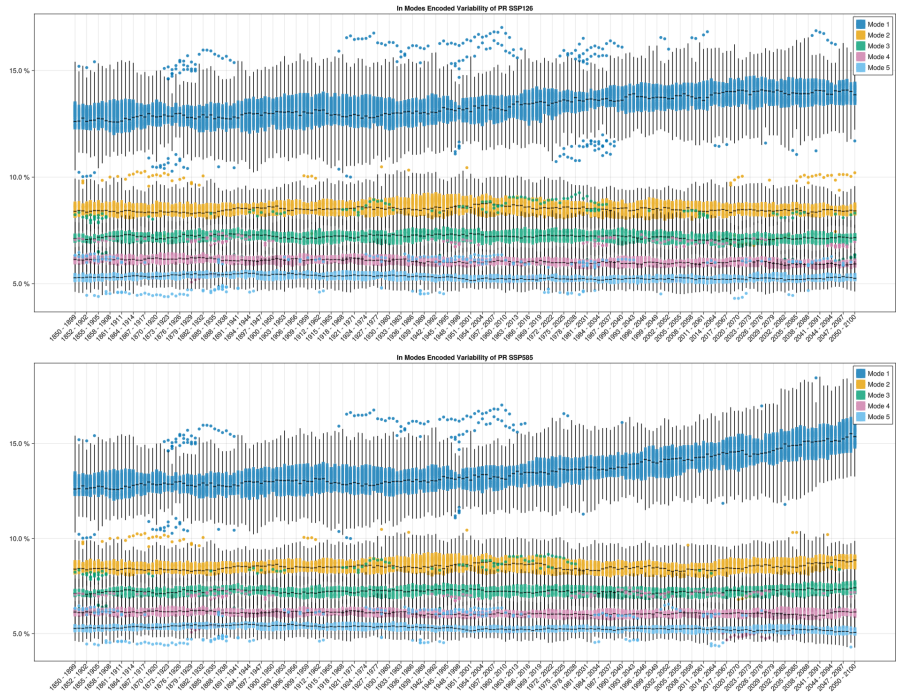

Figure 6.3: Same as Figure 6.1 but with precipitation

The comparison of mode variability evolution of precipitation EOFs (Figure 6.3) shows no significant changes of modes 3,4, and 5 between both evaluated scenarios. 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 encoded variability of around 8.5%. The primary EOF on the other shows significant differences across scenarios: While it has a far greater variability across members then the other modes and follows a general upwards 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.

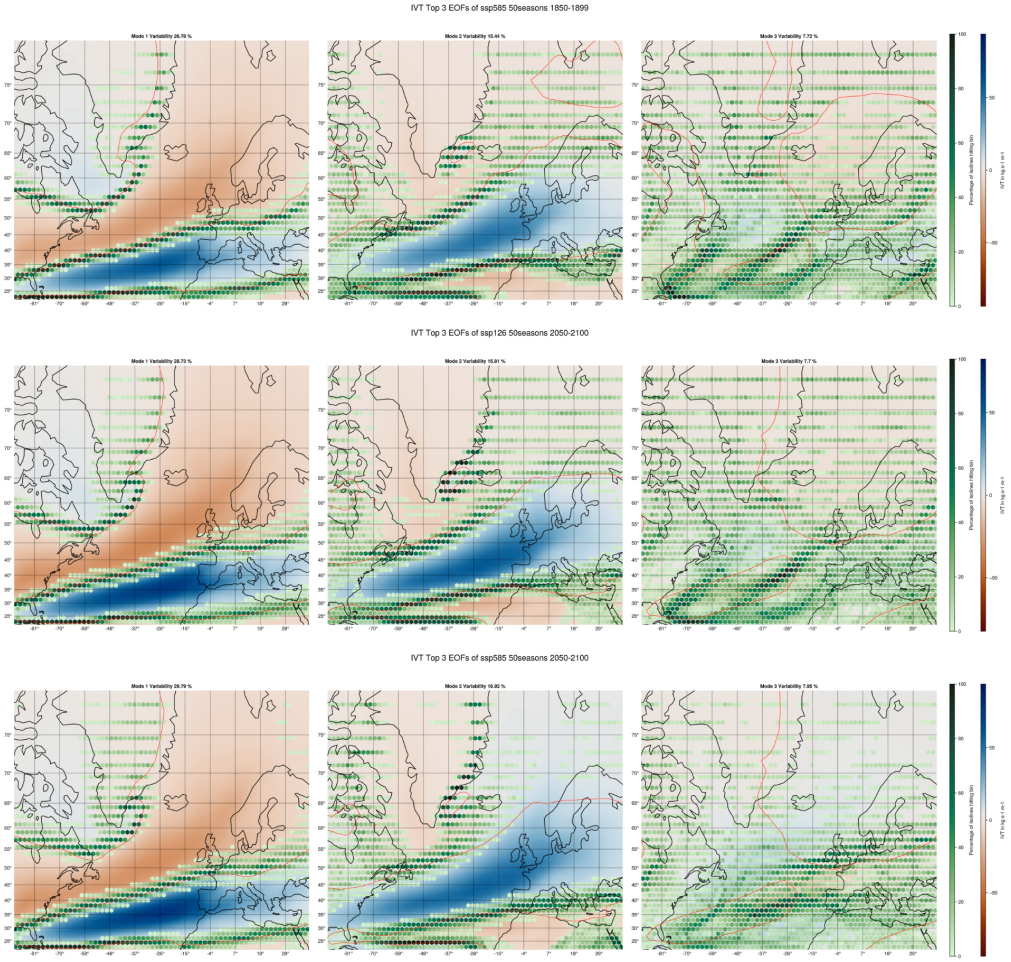

Figure 6.4: The top three EOFs of IVT data, with a 50 winter scope and hexbins visualizing the variability introduced by simulation members. The top row displays the state in the historical simulation (second half of 19th century), while middle (SSP126) and bottom (SSP585) display the state in the second half of the 21st century. The red line shows the contour line of zero of the preindustrial control simulation.

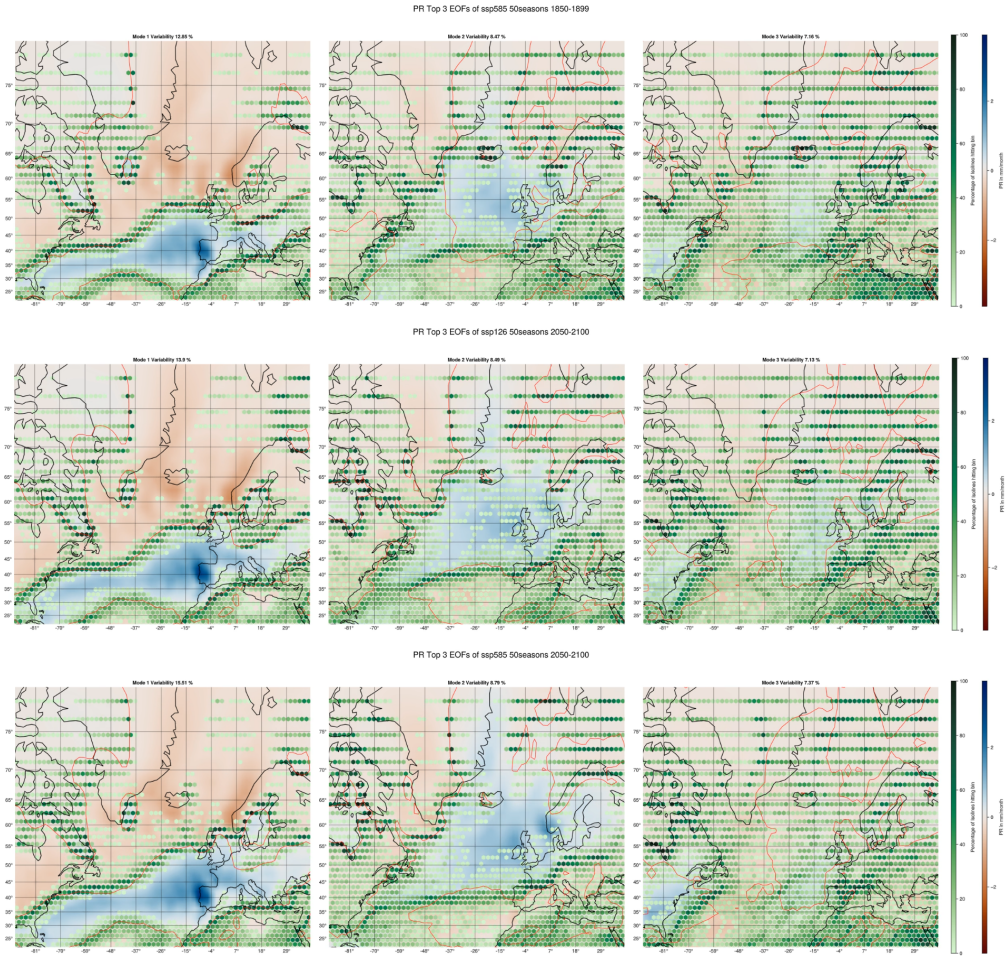

Figure 6.5: Same as Figure 6.4, but with precipitation data.

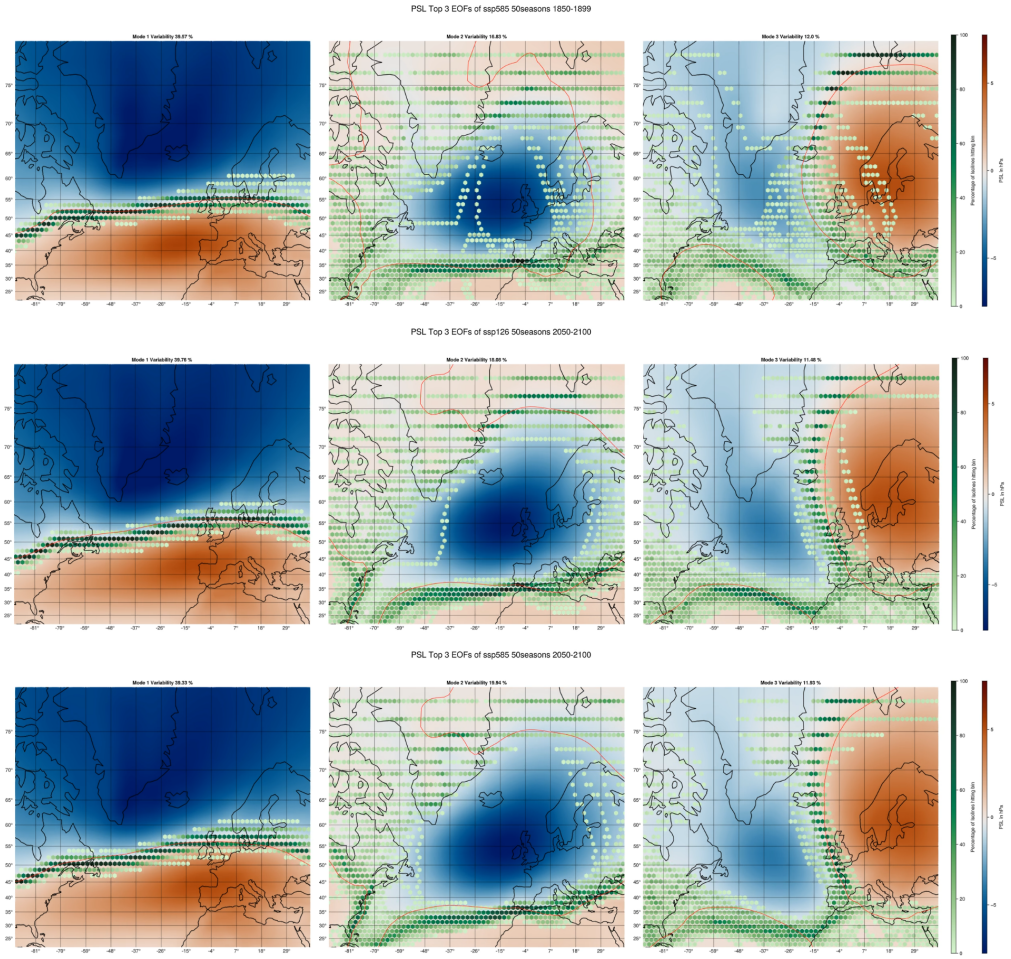

Figure 6.6: Same as Figure 6.4, but with sea level pressure data.

7 Conclusions and Future Work

7.1 Conclusions

7.2 Future Work

Bibliography

1. 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)". en. *Theoretical and Applied Climatology* 147:3-4, 2022, pp. 985–1002. ISSN: 0177-798X, 1434-4483. DOI: 10.1007/s00704-021-03864-x.

2. 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". en. *Monthly Weather Review* 134:4, 2006, pp. 1063–1080. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/MWR3123.1.

3. A. Barth. *NCDatasets.jl: a Julia package for manipulating netCDF data sets*. Issue: 97 Pages: 6504 Publication Title: Journal of Open Source Software Volume: 9 originaldate: 2017-06-25T20:42:08Z. 2024. DOI: 10.21105/joss.06504.

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

5. M. Böttinger and D. D. Kasang. *The SSP Scenarios*. en. Page.

6. G. Buckley. *Choosing good chunk sizes in Dask*.

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

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

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

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

11. S. Eckermann. "Hybrid σ–p Coordinate Choices for a Global Model". EN. *Monthly Weather Review* 137:1, 2009. Publisher: American Meteorological Society Section: Monthly Weather Review, pp. 224–245. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/2008MWR2537.1.

12. 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". en. *Journal of Geophysical Research: Atmospheres* 121:2, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD023379, pp. 931–948. ISSN: 2169-8996. DOI: 10.1002/2015JD023379.

13. 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". en. *Geoscientific Model Development* 9:5, 2016, pp. 1937–1958. ISSN: 1991-9603. DOI: 10.5194/gmd-9-1937-2016.

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

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

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

17. 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.

18. 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, 2020, p. 100254. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2020.100254.

19. E. Ghaderpour. "Map Projection", 2014.

20. L. Gimeno, R. Nieto, M. Vázquez, and D. Lavers. “Atmospheric rivers: a mini-review”. *Frontiers in Earth Science* 2, 2014. ISSN: 2296-6463.

21. 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”. en. *Geophysical Research Letters* 45:22, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL079249, pp. 12, 576–12, 584. ISSN: 1944-8007. DOI: 10.1029/2018GL079249.

22. A. Hannachi. “A Primer for EOF Analysis of Climate Data”. en.

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

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

25. J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. “An overview of the North Atlantic Oscillation”. en. 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.

26. 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, 2023. ISBN: 978-1-00-915789-6. DOI: 10.1017/9781009157896.

27. *JuliaIO/JLD2.jl*. original-date: 2015-07-02T21:59:50Z. 2024.

28. 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, 2022. DOI: 10.1002/jnm.3063.

29. 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”. en. *Journal of Visualization* 24:5, 2021, pp. 861–890. ISSN: 1343-8875, 1875-8975. DOI: 10.1007/s12650-021-00755-1.

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

31. 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.

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

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

34. 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, 2018, pp. 50–58. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2018.03.015.

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

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

37. 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". en. *Journal of Hydrometeorology* 9:1, 2008, pp. 22–47. ISSN: 1525-7541, 1525-755X. DOI: 10.1175/2007JHM855.1.

38. NOAA. *What's the difference between climate and weather? | National Oceanic and Atmospheric Administration*. en.

39. G. R. North, T. L. Bell, R. F. Cahalan, and F. J. Moeng. "Sampling Errors in the Estimation of Empirical Orthogonal Functions". EN. *Monthly Weather Review* 110:7, 1982. Publisher: American Meteorological Society Section: Monthly Weather Review, pp. 699–706. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.

40. 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". en. *Geoscientific Model Development* 9:9, 2016, pp. 3461–3482. ISSN: 1991-9603. DOI: 10.5194/gmd-9-3461-2016.

41. 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*. en. preprint. Preprints, 2023. DOI: 10.22541/essoar.168319746.64037439/v1.

42. K. Poethkow, C. Petz, and H.-C. Hege. "APPROXIMATE LEVEL-CROSSING PROBABILITIES FOR INTERACTIVE VISUALIZATION OF UNCERTAIN ISOCONTOURS". en. *International Journal for Uncertainty Quantification* 3:2, 2013, pp. 101–117. ISSN: 2152-5080. DOI: 10.1615/Int.J.UncertaintyQuantification.2012003958.

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

44. 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". en. *Earth System Dynamics* 7:2, 2016, pp. 371–384. ISSN: 2190-4987. DOI: 10.5194/esd-7-371-2016.

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

46. 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. Ro-

gelj, 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". en. *Global Environmental Change* 42, 2017, pp. 153–168. ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2016.05.009.

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

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

49. D. A. Salstein, R. D. Rosen, and J. P. Peixoto. "Modes of Variability in Annual Hemispheric Water Vapor and Transport Fields". en. *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.

50. 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". en. *IEEE Transactions on Visualization and Computer Graphics* 16:6, 2010, pp. 1421–1430. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.181.

51. J. S. Sawyer. "Man-made carbon dioxide and the "greenhouse" effect". *Nature* 239:5366, 1972, pp. 23–26.

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

53. U. Schulzweida. *CDO - Climate Data Operators*. Accessed: 2024-06-06. 2024.

54. 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". en. *Journal of Climate* 33:16, 2020, pp. 7179–7196. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-20-0011.1.

55. 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". en. *Journal of Climate* 33:1, 2020, pp. 263–279. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-19-0348.1.

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

57. A. C. Telea. *Data visualization: principles and practice*. cRC Press, 2014.

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

59. D. Vietinghoff. "Critical Points of Uncertain Scalar Fields", 2024.

60. 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". en. In: *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.

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

62. 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.

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

64. 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". en. *Atmosphere* 13:10, 2022, p. 1694. ISSN: 2073-4433. DOI: 10.3390/atmos13101694.

65. 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". en. *Journal of Geophysical Research: Atmospheres* 118:4, 2013, pp. 1585–1600. ISSN: 2169-897X, 2169-8996. DOI: 10.1002/jgrd.50089.

66. 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". en. *Geophysical Research Letters* 48:5, 2021, e2020GL091441. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2020GL091441.

67. T.-J. Zhou and R.-C. Yu. "Atmospheric water vapor transport associated with typical anomalous summer rainfall patterns in China". en. *Journal of Geophysical Research: Atmospheres* 110:D8, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2004JD005413. ISSN: 2156-2202. DOI: 10.1029/2004JD005413.

68. Y. Zhu and R. E. Newell. "A Proposed Algorithm for Moisture Fluxes from Atmospheric Rivers". en. *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.

69. 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". en. *Atmosphere* 11:7, 2020, p. 758. ISSN: 2073-4433. DOI: 10.3390/atmos11070758.

70. 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". en. *International Journal of Climatology* 38:6, 2018, pp. 2672–2685. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5452.