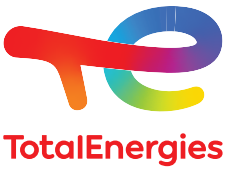
POLITECNICO DI TORINO
EURECOM INSTITUTE OF SOPHIA-ANTIPOLIS

Double Master of Science's Degree in
ICT for Smart Societies Engineering
Data Science and Engineering



Machine Learning applied to weather-related variables for energy trading activities

Supervisors

Prof. Maurizio FILIPPONE

Prof. Enrico MAGLI

Dr. Robin LOCATELLI

Candidate

Alessandro BALDO

2021/2022

Summary

Weather Regimes are large-scale, quasi-stationary and recurrent meteorological configurations. Their dynamics largely impact the main weather (temperature, precipitation, wind speed) and energy-related (power demand, hydro/wind/solar generation) variables. In the context of the forecasts, weather regimes are generally more predictable than usual 15-day forecasts of singular weather quantities. Hence, quantifying the consequences these conformations have on energy variables entails an opportunity to detect trends with larger anticipation, and it can be a valuable information for the energy trading sector. In our study, we thus give an overview of the physical nature of such regimes, presenting a set of advanced statistical methodologies in the Unsupervised Learning setting, in order to recognize and discern them, proving to outperform previous studies in literature. We leverage on such results to perform a fine-grained quantification on energy variables, envisioning pervasive applications in the trading of energy commodities. Finally, we provide a forecasting method projecting the results collected in the previous steps into a sub-seasonal window (until 6 weeks in advance), which we apply on the Winter 2021-2022.

Summary

Les régimes de temps sont des configurations météorologiques à grandes échelles, quasi-stationnaires et récurrentes. Les variables météos usuelles (température, précipitation et vitesse de vent) et leurs équivalents dans le domaine énergétique (demande électrique, production hydraulique/solaire/éolienne) sont fortement corrélées avec ces régimes. De plus, les météorologues s'intéressent beaucoup à ces configurations car les erreurs de prévisions sont généralement plus petites lorsqu'on s'intéresse à des structures de grandes échelles. Ainsi, quantifier précisément les impacts de ces régimes sur les variables énergétiques donne l'opportunité de détecter des tendances météorologiques avec un bon degré d'anticipation, ce qui peut apporter des informations essentielles aux traders de l'énergie. Dans notre étude, nous caractérisons la structure physique des régimes de temps et nous décrivons également des méthodes statistiques avancées dans le domaine de Unsupervised Learning pour détecter ces structures. En appliquant ces méthodes à un jeu de données historiques des conditions climatiques, nous avons été capables de quantifier les impacts des régimes de temps sur plusieurs variables énergétiques pour ensuite les appliquer au trading de l'énergie. Finalement, une méthode de prévision des tendances a été développée à un horizon temporel de 6 semaines et appliquée pour le début de l'hiver 2021-2022.

Acknowledgements

For whoever it may be concerned with the following acknowledgements, I decided on purpose to omit each explicit name of any of the friends involved, hoping that many people could recognize themselves in some of the words, visualizing them with memories, events and feelings we shared along the years.
I want to say thanks to *you*, to have helped me in never settling with my achievements, rather always believing in me, to push me beyond my capabilities, and asking myself more and more. Thanks for having enlightened me, for having revived the fire of my curiosity in periods when I could not have managed it by myself.
Also, thanks to *you* that, in our ups and downs due to my sometimes difficult personality, always stayed along my side, giving me the personal support I needed during my darkest moments, however never pretending to have anything in return.
Thanks to *you*, constantly and historically with me. Thanks to have taught me to be resilient, to accept my weaknesses, and to simply, but invaluably, be present whatever the circumstance was.
Thanks to the Politecnico di Torino, in my beloved Turin. The years spent in its corridors and classrooms are a memory I will always bring with myself, and which I will always look back with nostalgia.
Thanks to Eurecom that, although the pandemic divided us, opened up new perspectives in me, challenging me outside my personal comfort zones, and clarifying my long-term objectives. Thanks to have given me the possibility to live a sparkling setting, with the most talented students, now friends, further motivating me to excel, and to set unprecedented ambitious goals.
Thanks to TotalEnergies, and in particular to all the friends of the Market Research Analysis team: you made these months unforgettable. A special thanks to my supervisor, mentor and friend, who has guided me in these months, and to my manager: a person of value, without whom this experience would not have been possible.
Finally, the most important acknowledgement goes to my family: *Mamma*, my first motivator and supporter; *Papà*, for having me taught to stay grounded, and to be self-critical; my grandparents, the ones still here with me, and the ones already foreseeing this moment long ago, to whom I dedicate this achievement.

Table of Contents

List of Tables VIII

List of Figures IX

Acronyms XIII

1 Introduction 1
- 1.1 Introduction to the study . . . 1
- 1.2 Energy Markets . . . 2
- 1.3 Outline on weather variables . . . 6
  - 1.3.1 Geopotential Height . . . 6

2 Data 10
- 2.1 Meteorological Data . . . 10
  - 2.1.1 Historical Data: ERA5 and Energy Variables . . . 11
  - 2.1.2 Weather Regimes - Météo-France . . . 12
  - 2.1.3 Sub-seasonal Forecasts . . . 13
- 2.2 Data Pre-processing . . . 14
  - 2.2.1 Empirical Orthogonal Function (EOF) Analysis . . . 15
  - 2.2.2 Principal Component Analysis (PCA) . . . 18
  - 2.2.3 Variational Autoencoder (VAE) . . . 18

3 Models 25
- 3.1 K-Means . . . 26
- 3.2 Mixture Models . . . 27
  - 3.2.1 Gaussian Mixture Model (GMM) . . . 29
  - 3.2.2 Dirichlet Process Mixture Model (DPMM) . . . 30
- 3.3 Results . . . 33
  - 3.3.1 Theoretical Results . . . 35
  - 3.3.2 Comparison with Météo-France . . . 41
  - 3.3.3 From Statistical To Physical Properties . . . 43

3.3.4 Transitions 45

4 Quantitative Analysis in the Energy domain 46
4.1 Energy-related variables 46
4.2 Quantifications 48

5 Sub-seasonal Forecasts 53
5.1 Case Study: when the forecast is accurate 54
5.2 Case Study: when the forecast is inaccurate 55

6 Conclusions 59

Bibliography 61

List of Tables

2.1 Set of energy variables coming from ERA-5 dataset and TSO *variables not available for all the considered countries 12

3.1 Evaluation Metrics for each model. ↑/↓ indicates that the score has to be maximized/minimized 35
3.2 Best performing models, under the two dimensionality reductions. ↑/↓ indicates that the score has to be maximized/minimized. The column “Objective" indicates which score was optimized to in the Cross-Validation process of the model 36
3.3 Frequencies of the regimes as predicted by (Cassou et al. 2011; van der Wiel et al. 2019a; Luo et al. 2012; Meteo-France n.d.) vs. Frequencies as predicted by K-Means under PCA and VAE reduction schemes. We notice how K-Means is highly sensitive to the inferred mapping of VAE 37
3.4 Frequencies of the regimes as predicted by (Cassou et al. 2011; van der Wiel et al. 2019a; Luo et al. 2012; Meteo-France n.d.) vs. Frequencies as predicted by Mixture Models under PCA and VAE reduction schemes. We notice how under PCA statistics tend to be more misaligned 40
3.5 Frequencies (%) of the intermediate transitions between pairs of regimes, as predicted by our Mixture Models. 40
3.6 KL-divergences of the regimes' distributions in GMM (left) and DPMM (right) 44
3.7 Transition Probabilities as predicted by GMM (left) and DPMM (right) 45

List of Figures

1.1 Snapshot of all the reference prices for LNG (Liquefied Natural Gas) energy market in Europe. The TTF (Netherlands) is the reference price, both for international and internal (whether the other prices are expressed in they form TTF + Δ trades 4

1.2 Illustration of the establishment of the electricity price, based on the actual supply/demand and the productive expectation of the power plants. Power plants are dispatched according to the merit order, based on the related short-run costs 5

1.3 Geo-potential maps of daily observation, normal and anomaly . . . 6

1.4 Visualization of the four weather regimes (Weather et al. n.d.) in the North-Atlantic and European zone. From left to right, up to down: NAO+, presenting high pressure in the mid-latitude level, and low pressure over the sub-arctic zone; Scandinavian Blocking (SB), characterized by high pressure over the Scandinavian peninsula; Atlantic Ridge (AR) configuration presenting higher pressure over the Atlantic Ocean region; NAO-, reporting general low pressure over Europe and high pressure in the sub-arctic region 8

2.1 Operational and Ensemble models in a 15-days forecast scenario. In **black** operational model, in red the mean of the ensemble. It is worth to notice the increasing volatility of the members of the ensemble for further forecasted dates 11

2.2 Visualization of the four regimes in the North-Atlantic/European zone, as predicted by Météo-France (Meteo-France n.d.). It is possible to identify: AR (top left); NAO+ (top right); NAO- (bottom left); SB (bottom right) 13

2.3 Visualization of the 45-days sub-seasonal forecast provided by ECMWF (Ferranti et al. 2011) 14

2.4 Visualization of the four main EOFs (Empirical Orthogonal Functions) in the North-Atlantic/European zone. It is possible to identify: a NAO+ resembling component (top left); an anti-Scandinavian Block SB- (top right); an anti-Atlantic Ridge AR- component (bottom left) 16
2.5 Bar Visualization (left) of the main Principal Component (PC-1) of the Time Series, in the North-Atlantic/European zone, grouped by the January-February-December months of the same solar year. PC is here standardized to have zero mean and unit variance. When the values are out of the range ±1, they are generally associated to more extreme meteorological events. Associated to it, the NAO Index (Climate Prediction Center n.d.) evolution during winter months: it is evident the high correlation with PC-1. On the right, a density visualization of the first two PCs 17
2.6 Visualization of an Autoencoder architecture(Elbattah et al. 2021) . 19
2.7 Visualization of a Variational Autoencoder (VAE) architecture(LearnOpenCV n.d.) 19
2.8 t-SNE applied on the PCA-reduced (top row) and VAE-reduced (bottom row) spaces at different levels of perplxities: the VAE-encoded features leads to a better separability of the clusters, as further proved by the better minimized KL-divergence objective . . 23

3.1 K-Means with 4 clusters. AR (top left), NAO+ (top right), NAO- (bottom left), SB (bottom right) 37
3.2 K-Means with 7 clusters 38
3.3 Mixture Models with 4 clusters. AR (top left), NAO+ (top right), NAO- (bottom left), SB (bottom right) 39
3.4 Variational Dirichlet Process Mixture Model with 7 clusters 41
3.5 Confusion matrices between our 4 cluster models and Metéo-France historical predictions. The Mixture Models are inline with the agency's predictions. K-Means algorithm produces better scores, since it is aligned with the modeling choice of Meteo-France. From this illustration, it is evidenced how each model mostly strives with the Atlantic Ridge (AR) and Scandinavian Blocking (SB) configurations 42
3.6 ROC curves of the three models: K-Means with PCA reduction (left), DPMM (center) and GMM (right) with VAE reduction 42
3.7 Historical counts of winter-days by regime 43
3.8 Visualization of the distributions of the clusters of the two mixture models along the two main modes of variability 44

4.1 Examples of relationships across weather and energy-related variables 47

4.2 Wind Load Factor Deviation (%) (left), Solar Load Factor Deviation (%) (center), Power Demand (GW) during business days (right) under the different regimes for EU-7 countries 48
4.3 Geographic visualization of the regimes and countries relationships of the Wind Load Factor and Power Demand variables 49
4.4 Example of historical distributions of the Wind and Solar Load Factor energy variables, isolated to an individual country (Germany) 50
4.5 Example of monthly historical distributions of the Wind and Solar Load Factor energy variables, isolated to an individual country (Germany) 51
4.6 Occurrence (%) of extreme events in EU-7 countries in the four different regimes for Wind (left) and Power Demand (right) 52

5.1 Average weights of the sub-seasonal forecasts of ECMWF by regime 54
5.3 Quantification of the Sub-seasonal forecast on the Wind Load Factor Anomaly in Germany 56
5.5 Quantification of the Sub-seasonal forecast on the Wind Load Factor Anomaly in Spain 58

Acronyms

ML
Machine Learning

GMM
Gaussian Mixture Model

DPMM
Dirichlet Process Mixture Model

DP
Dirichlet Process

gpH
Geopotential Height

hPa
hecto-Pascal

NAO
North Atlantic Oscillation

SB
Scandinavian Blocking

AR
Atlantic Ridge

EOF
Empirical Orthogonal Function

PCA

Principal Component Analysis

AE

AutoEncoder

VAE

Variational AutoEncoder

S2S

Sub-seasonal

DJF

December-January-February

TSO

Transmission System Operator

OTC

Over-the-Counter

LNG

Liquefied Natural Gas

EEX

European Energy Exchange

ICE

Intercontinental Exchange

CME

Chicago Market Exchange

SRMC

Short-Run Marginal Cost

Chapter 1

Introduction

1.1 Introduction to the study

Modeling weather dynamics requires a thorough understanding of the field and its underlying physical processes. The chaotic nature of these systems entails the main obstacle to accurately map the evolution of meteorological configurations, especially when considering small-scale events.
Looking instead at large-scale patterns allows to enlarge the forecasting window beyond some few days, and it could be also useful to track phenomena at the finer-grained level. In this context, weather regimes still represent a partially explored branch of meteorology, while being some of the main drivers and source of variability in the climatic conditions.
Despite their recurrent and quasi-stationary nature, the innermost difficulty is to properly recognize them, and, consequently, quantifying the impact of each of these physical configurations.
Application domains beyond meteorology are pervasive in daily life scenarios, among which the energy sector stands out, with a particular emphasis on energy trading activities. To render the idea, some of the main influenced factors are:

- production: this is especially true when dealing with renewable energy systems, like solar, wind power and hydro-electric productions
- transport: many goods are still transported over rivers, such as LNG (Liquid Natural Gas). The monitoring of water-levels is thus central, being highly correlated with meteorological phenomena like precipitations, snow melting, temperatures etc.

and, more generally

- the demand and consumptions of end-customers, influencing the *base* and

peak loads, and thus the necessity for an energy producer/seller/provider to intervene on the market

Hence, the possibility to accurately predict (up to a certain extent) meteorological patterns represents a possibility to build weather-driven financial strategies with some anticipation.
In this context, systems of physical equations, while central in tracking the short-term dynamics, are of limited effectiveness, being strongly sensitive to the initialization conditions. Therefore, we recur to the means of statistical learning techniques, as a trade-off offering cheaper computation, better performances, while preserving an overall higher interpretability of the dynamics. Also, we extend our focus on the medium-term range (sub-seasonal), possibly entailing a "window of opportunity" for energy trading applications we aim to encompass. Thereby, we are interested in giving a proper quantification of energy variables, strictly dependent on the regimes recognition task. Finally, we leverage on this knowledge to perform sub-seasonal forecasts, trying to catch signal up to 40 days with anticipation.
Our study is mostly focused on the North-Atlantic European region, during the winter period. However, the pipeline we develop is highly portable, and easily adaptable to different input data typologies, as well as to different geographic zones and seasons.
The dissertation is organized as follows: in Section 1.2 an introduction to energy markets is presented, together with the main financial instruments adopted in the setting; Section 1.3 outlines the most relevant energy variables employed in the project, and firstly introduces the concept of *weather regimes*. Chapter 2 gives an overview of the main datasets used across the study, as well as presenting the data pre-processing techniques conceived for the project (Section 2.2). Subsequently, Chapter 3 covers an extensive dissertation on the Machine Learning algorithms used to cluster weather regimes, as well as presenting the experimental campaign from Section 3.3 on. Finally Chapters 4 and 5 integrate the two final steps of the pipeline, respectively discussing the statistical assessment and forecasting procedures we have performed in the energy domain.

1.2 Energy Markets

Energy products are located inside the more general category of commodities, i.e. any sort of goods that have an effective physical and tangible realization, outside the logics of the market. Trading on commodities could then happen either in a physical or speculative way, whether or not the delivery of the underlying good has its effective realization.
The nature of the commodities' market makes it singular, involving long-term planning and many factors at the support of the decision-making process. As

a consequence, the main financial instruments are often tailored to this context, including:

- *futures* contracts: a well-established financial artifact, representing an *obligation* of the subscriber to obtain an agreed quantity of a commodity at a certain delivery date, for a price reflecting the financial state of the good at the stipulation date. There exist several variants of these contracts, at this point covering any major commodity, in any market and for periods up to tens of years forward in the future.
  Of course, given that these contracts are mostly closed before their expiration (i.e. the physical delivery of the good), they often represent a dynamic instrument for speculation

- *forwards* contracts: fundamentally relying on the same concept as for the futures, but instead of being traded on an Exchange, they are traded *over-the-counter* (OTC). Hence, the broker acts outside a regulated market, allowing for more customizable contracts

- *options*: they are conceptually similar to futures, but the subscriber here pays a *possibility* to buy/sell a commodity at a certain date, and at a certain price (still, reflecting the present situation). Again, they are a further speculative strategy, often used for hedging purposes (i.e. to reduce the risk associated to other concurrent open trades)

Energy markets typically relies on regulated entities, called *exchanges*. By definition [1], an Exchange is an economic system inside which goods and services are produced, distributed, and exchanged by the forces of price, supply, and demand. At a worldwide level, the major exchanges for traded volumes are represented by:

- *EEX* (European Energy Exchange): the European reference market, located in Germany and interconnecting the markets of energy and linked products. It represents a very liquid market, born as a response to the previous well-established framework in Europe, composed of several monopolies, characterizing the member countries.
  The main commodities are represented by Natural Gas (NG) and Power. Inside this exchange each country has its own reference price (Figure 1.1), however the *TTF* price is typically adopted as the main reference, with the other prices often expressed as a function of it (i.e., TTF $\pm \Delta$)

- *ICE* (Intercontinental Exchange): an American system, whose main products are futures, energy commodities and OTC derivatives

[1] https://sociologydictionary.org/market-exchange

- *CME* (Chicago Market Exchange): again, a USA-based system, offering a wide variety of futures

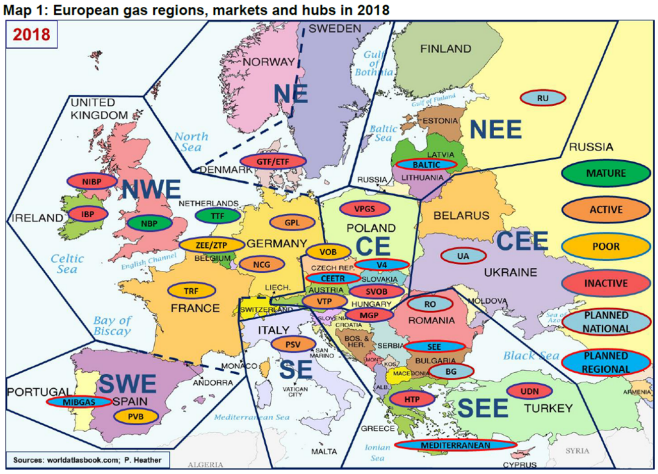


Figure 1.1: Snapshot of all the reference prices for LNG (Liquefied Natural Gas) energy market in Europe. The TTF (Netherlands) is the reference price, both for international and internal (whether the other prices are expressed in they form TTF $+ \Delta$ trades

In case of power markets, electricity prices are based on a supply-demand equilibrium. Power plants are dispatched according to the *merit order*, from the lowest to the highest short-run costs, until reaching the level of the demand.
The last unit dispatched in the merit order is called the *marginal unit*. Players in the market adopt an economic behavior:

- Generators produce electricity if the compensation is higher than the production costs
- Generating companies build new plants if they are profitable (i.e. if price signals are high enough)
- Consumers buy at the lowest possible price

In other words, the market is based on a supply-demand equilibrium, whose each point corresponds to a market price (Figure 1.2). An equilibrium point is reached at each period of time. We define then as *Short-Run Marginal Cost* (SRMC) the average of the variable costs of the marginal dispatched units, over the year. When capacity is sufficient, the calendar power price should be equal to it. Price cannot be neither higher (otherwise the extra-margins from selling at higher prices could be challenged by competitors willing to settle for less), nor lower (in this case,

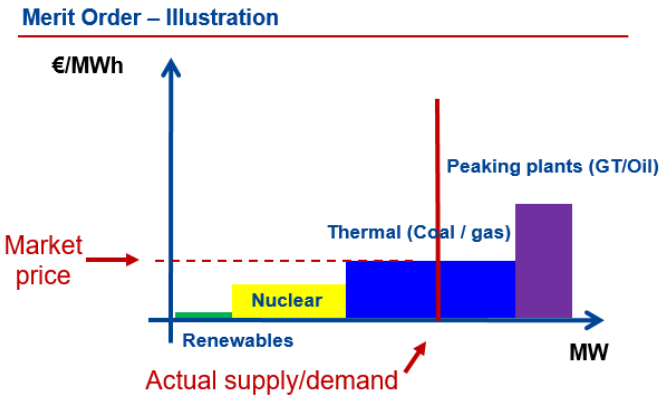


Figure 1.2: Illustration of the establishment of the electricity price, based on the actual supply/demand and the productive expectation of the power plants.
Power plants are dispatched according to the merit order, based on the related short-run costs

producers would record losses if selling below it).
In case of capacity shortage, a scarcity rent appears during some hours, i.e. a mark-up over the SRMC compensating for the additional investments and fixed costs of a new marginal unit. Depending on the market structure, this mark-up may either be reflected implicitly in the energy market price, or as an explicit capacity price.
In this situation, we can have then two types of movements of the merit order:

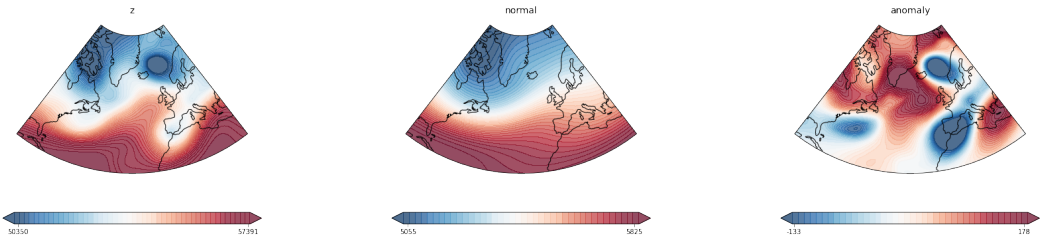
- horizontal movements, due to technical outages and/or unavailabilities and/or scarce renewable energy production
- vertical movement, due to commodity price fluctuations

For our study, we are especially interested in the first typology of movements, especially when quantifying how a particular weather configuration can affect fluctuations in this sense, both on the production (e.g. higher than average wind generation compensating for the demand) or on the power demand (for example, higher temperatures leading to lower power demand).

1.3 Outline on weather variables

1.3.1 Geopotential Height

The primary part of our study focuses on the establishment of the so-called *weather regimes*: meteorological configurations at a regional/sub-continental level, defining a proper atmospheric state, and highly influencing climate phenomena such as precipitations, winds, water levels, temperatures etc.



(a) Observation of geo-potential height at 500hPa (Z500)

(b) 30-years “flat" normal of geo-potential height at 500hPa (Z500).

(c) Anomaly of geo-potential height at 500hPa (Z500)

Figure 1.3: Geo-potential maps of daily observation, normal and anomaly

They are usually addressed considering the *geopotential height* (gpH) or Sea Level Pressure (SLP) variable: the altitude at which a constant atmospheric pressure level is measured (i.e. the altitude of an iso-pressure atmospheric level curve). Normally, the reference pressure value is Z500, corresponding to 500 hPa (hecto-Pascal), and taking into account those physical phenomena occurring in the mid/low Stratosphere, thus majorly driving climate events at the surface level. However, across the studies other pressure levels have shown to be suitable to accomplish the task (Hertig et al. 2014; Vautard 1990; Meteo-France n.d.; Vrac et al. 2007).
Instead of considering the raw gpH observation, the so-called *anomaly* (Figure 1.3c) is examined. It is a physical quantity addressing the variation with respect to the *normal* behavior (Figure 1.3b).
We define the normal as corresponding to an average involving (typically) the last 30 years of observations of the considered period of the year, thus resembling the mean expected behavior. It is a position-wise quantity, that, in the case of the geopotential height, shows a clear trend: it increases around the tropics (typically characterized by higher-pressure regimes), and decreases moving towards extreme latitudes.
Evaluating it often requires some processing, in order to eliminate any sort of seasonality that could be present when considering periods spanning across different seasons. Further, it has proved to highly influence the sensitivity of algorithms in the modeling pipeline, especially when leading to large anomaly values. For this reason, in our project we consider two types of normal: a "flat" normal, independent of the time, and a more "dynamic" normal, varying at different temporal granularity levels (month, week, and day).

Pressure zones are generally divided between low and high pressure zones, consequently corresponding to low and high values of geopotential height. In a 3D coordinate systems, they can be mathematically conceived as local minima/maxima on the isobaric plane. Under a meteorological point of view, low pressure zones are usually drivers of precipitations and unstable climatic conditions. Conversely, high pressure areas tend to be drier and with more stable setups. The geographical zone between an high pressure and a low pressure clusters experiences instead a general windy climate, the more accentuated, the closer the two clusters are. Finally, according to the relative positions of the two pressure masses, wind currents take either clockwise or anti-clockwise directions.
Thereby, weather regimes are recurrent, and quasi-persistent spatial distributions of these pressure masses, presenting a well-defined relative location of positive and negative pressure zones. They vary across geographical areas (Robertson et al. 1999; Bruyère et al. 2016), and between different seasons, being more prominent during winter and summer (Cassou et al. 2005).

Our project is mainly centered on the North-Atlantic and European zone during

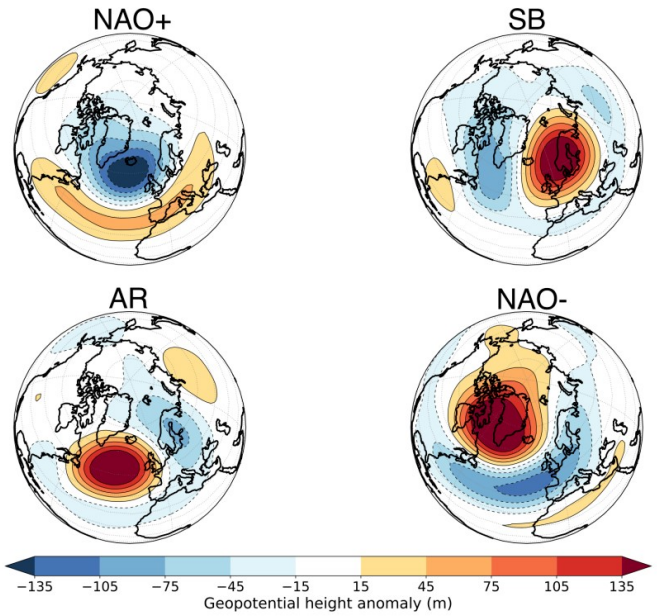


Figure 1.4: Visualization of the four weather regimes (Weather et al. n.d.) in the North-Atlantic and European zone. From left to right, up to down: NAO+, presenting high pressure in the mid-latitude level, and low pressure over the sub-arctic zone; Scandinavian Blocking (SB), characterized by high pressure over the Scandinavian peninsula; Atlantic Ridge (AR) configuration presenting higher pressure over the Atlantic Ocean region; NAO-, reporting general low pressure over Europe and high pressure in the sub-arctic region

the winter season. In such context, the four theoretical regimes (Figure 1.4) (Straus et al. 2007; Grams et al. 2017; Falkena et al. 2020; Cassou et al. 2005; Cassou et al. 2011; van der Wiel et al. 2019a) are:

- *NAO positive* and *NAO negative* (NAO+ and NAO-, respectively). North Atlantic Oscillations represent the two main weather configurations, respectively localizing high pressure (low pressure for NAO-) systems in the Mediterranean Europe/mid-Atlantic region, and low pressure (respectively, high pressure) in the northern Europe/Greenland.
  Under NAO+, winters are generally warmer over all Europe, with some precipitations localizing in the north-western zone of the continent. The Atlantic is instead more likely interested by storms.
  NAO- is characterized by colder winters in north-western Europe, and a wetter season in south-eastern Europe. Overall, there is less probability of winter storms.

- *Scandinavian Blocking* (SB). As the name recalls, the high pressure interests the Scandinavian peninsula and the northern Europe, while low pressure is mostly present over the Greenland and northern Atlantic America.
  Under SB, winters are typically dry and cold in the wester and central zones of Europe.

- *Atlantic Ridge* (AR). Again, as the name suggests, high pressure is over the entire Atlantic Ocean region, while low pressure is concentrated over the Scandinavia and, more limitedly, on the Europe.
  Under AR, temperatures are slightly below normal over Europe, and typically dry in the western part.

NAOs tend to be the pre-dominant regimes during winter, accounting for most of the variability. Due to this, we often resort to a custom indicator called *NAO Index* (Climate Prediction Center n.d.), whose positive values are associated to NAO+ persistances, while negative scores are more resembling NAO-. Further, its amplitude is often used to track extreme events, whether magnitudes are extreme.

Chapter 2

Data

2.1 Meteorological Data

Meteorological data are divided in raw and synthetic data. Raw meteorological data often comes in a temporal-spaced grid-based format. The density of the grid is a discriminatory feature, reflecting into the precision of mapping phenomena at a local-regional level.
Every cell in the grid (vertex of a cell) derives from a *data assimilation* process, which, starting from a set of initial weather conditions as measured by ground stations, satellites, planes, boats etc., allows to project them in a regular format, at a particular point in time, leveraging on systems of physical equations.
Meteorological forecasts by means of physical models fall into this category, and are divided in:

- Short-term or 15-days forecasts
- Medium-term or *sub-seasonal* forecasts: quite a recent concept, they analyze periods covering up to 6 weeks forward in the future
- Long-term or *seasonal* forecasts: adopted to get an overview over the next season

Dealing with chaotic systems, short-term forecasts are reliable just for few days after the forecasting date. This problem is typically addressed as "butterfly effect", where a minor change in the aforementioned initial conditions leads to potentially divergent forecasts in longer time horizons.
Synthetic data relies instead on the use of statistical models, typically inferring quantities from raw meteorological measures. For instance, this is the case for most of the forecasts of energy variables (e.g. wind/solar power generation derived respectively from wind speed/solar irradiance measures).

Further, synthetic data is often useful for data augmentation purposes, allowing to project data far in the past, where either raw measurements are missing or there was not the possibility to obtain them (e.g., a standard example is represented by renewable power generation when no renewable plant was still existing).
A further necessary differentiation refers to the two types of forecast models: *operational* and *ensemble* models. The former is a singular fine-tuned instance, usually representing the best estimation of the forecast. It is a deterministic model, leveraging on the best knowledge we have of the initial conditions, by combining and comparing several different measurement sources. It usually exploits the best resolution, both vertically and spatially, as well as the best physical parametrization of the underlying model. However, initial conditions are likely to differ from the ground truth. For this reason, ensemble models are introduced. A set of initial conditions is obtained by perturbing them with noise, thus allowing to explore the space of errors. Each of these "members" (i.e., a singular configuration of the ensemble) is plug into a model, typically on a lower resolution, due to computational constraints. Sometimes the model can be perturbed too, to increase the robustness. As a consequence, an ensemble of forecasts is obtained, which are typically used to get some insights into longer-term trends and patterns. Figure 2.1 reports an illustration of the difference between the two categories.

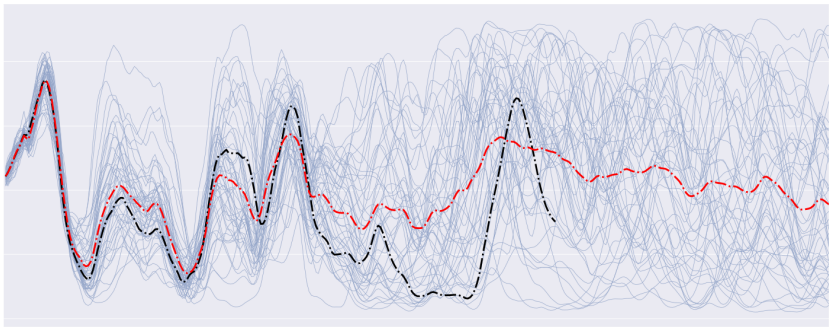


Figure 2.1: Operational and Ensemble models in a 15-days forecast scenario. In **black** operational model, in red the mean of the ensemble. It is worth to notice the increasing volatility of the members of the ensemble for further forecasted dates

In terms of data availability, forecasts are mostly proprietary measurements, coming in quite expensive solutions. For the scopes of this project, we only rely to open data, whether they are historical or raw re-analyzed collections.

2.1.1 Historical Data: ERA5 and Energy Variables

Our main data source of reference is Climate Data Store (Store n.d.[a]), a web platform exposing proper APIs and allowing to customize the data requests according

to the needs. It is a database containing re-analyzed raw data, offering a vast pool of well-noted meteorological models.
For the first part of the project, we use the standard *ERA5* dataset on single pressure-levels (Hersbach et al. 2020; Store n.d.[b]), collecting daily observations from 1979, up to the present.
The reference period we choose to examine is the winter-time, restricted to the December-January-February (DJF) months. In this way, we ensure to have similar conditions, avoiding any possible effect of seasonality.
Under the geographic point of view, we consider the North-Atlantic/European region (Cassou et al. 2011; Falkena et al. 2020; Grams et al. 2017; van der Wiel et al. 2019a), corresponding to an area of over 40 million km^2 (latitude in $[20°, 80°]$, longitude in $[-90°, 30°]$) and ensuring to capture the entire regional envelope of the weather regimes. The regular grid of the dataset is composed of 115′921 points, for a density of 0.25 lat-lon degrees (about 30 km). We place high importance on the high density characterizing the dataset, since it allows us to adopt more complex techniques than what the literature exposes.
For the quantification part of the dissertation, we take advantage of secondary weather variables coming from ERA-5, such as Temperature, Solar Irradiance, Sea Level Pressure, Wind Speed, as well as raw and synthetic energy data directly obtained from TSO (Transmission System Operator), a network of European operators providing the measurements on their grid (some examples are RTE in France, Terna in Italy). Table 2.1 reports the set of variables we take into account.

Energy Variables			
Wind Offshore Obs (GW)*	Wind Onshore Obs (GW)	Wind Capacity (GW)	Wind Load Factor
Solar Photo Obs (GW)	Solar Thermal Obs (GW)*	Solar Capacity (GW)	Solar Load Factor
Hydro Run of River Obs (GW)*	Hydro Reservoir Obs (GW)*	Hydro Capacity (GW)	Hydro Load Factor
Load			
Weather Variables			
Temperature (°C)	Precipitation (mm)	Solar Irradiance (W/m^2)	
Wind Speed - Sea Level (m/s)		Wind Speed - 100m (m/s)	

Table 2.1: Set of energy variables coming from ERA-5 dataset and TSO
*variables not available for all the considered countries

2.1.2 Weather Regimes - Météo-France

In modeling weather regimes, a definition of ground truth does not properly exist. This is evident across the numerous studies, where results tend to be slightly misaligned.
Weather regimes are phenomena which have been only recently adopted for high-level applications of side sectors, therefore freely available datasets tend to be scarce.

Météo-France (Meteo-France n.d.) is a national weather agency, which offers us the access to their historical collection of daily weather regimes from 1991 up to the present, as predicted by their ensemble centroid-based model. For our scopes, we decide to plug them into our a-posteriori validation framework, after we empirically validate their reliability.
As it is discussed in the next sections, we include in our results a probabilistic characterization of the ensemble results, although we specify how this is only determined by a pure frequentist assumption.
Figure 2.2 reports the centroids of the four weather regimes Météo-France predicts.

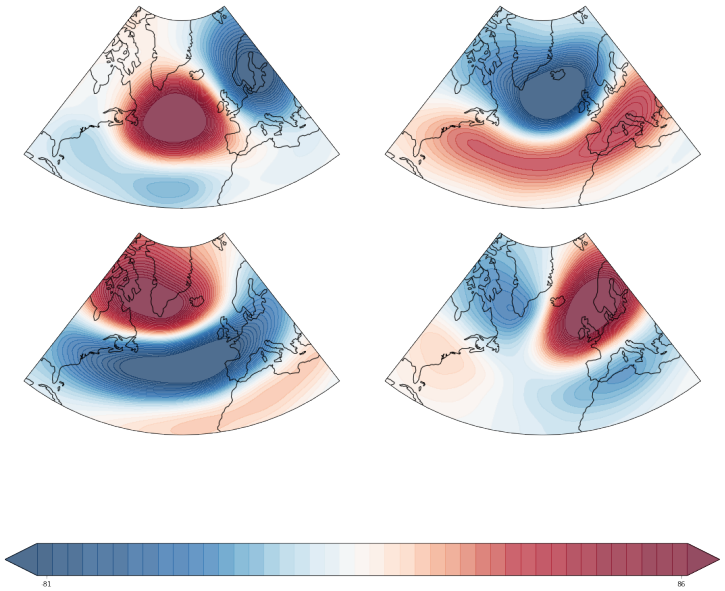


Figure 2.2: Visualization of the four regimes in the North-Atlantic/European zone, as predicted by Météo-France (Meteo-France n.d.). It is possible to identify: AR (top left); NAO+ (top right); NAO- (bottom left); SB (bottom right)

2.1.3 Sub-seasonal Forecasts

Sub-seasonal forecasts entail a new perspective in data-driven energy markets. Since large-scale configurations tend to be much more predictable and recurrent, medium-term forecasts represent an opportunity in several application domains, thus having recently led to a major development of sub-seasonal models.
Under a meteorological point of view, the stationarity (and thus the predictability) of a regime happens due to events occurring in the Stratosphere, and partially blocking the regime circulation. In our study, we aim at detecting some trends in

energy variables by correlating the forecasted weather regimes, with the historical quantification we perform on them.
Data for sub-seasonal forecasts are retrieved by the ECMWF (European Centre for Medium-Range Weather Forecasts) (Ferranti et al. 2011), releasing ensemble models' predictions twice a week. Again, in processing this collection we adopt a frequentist perspective to enable a probabilistic modeling on the problem.
In their modeling, they adopt a 5-clusters model (four standard clusters, and an additional "Unknown" cluster), which we easily reconduct to a canonical 4-clusters assumption, due to reasons which will be clear later on.
Figure 2.3 offers a snapshot on how sub-seasonal forecast looks like.

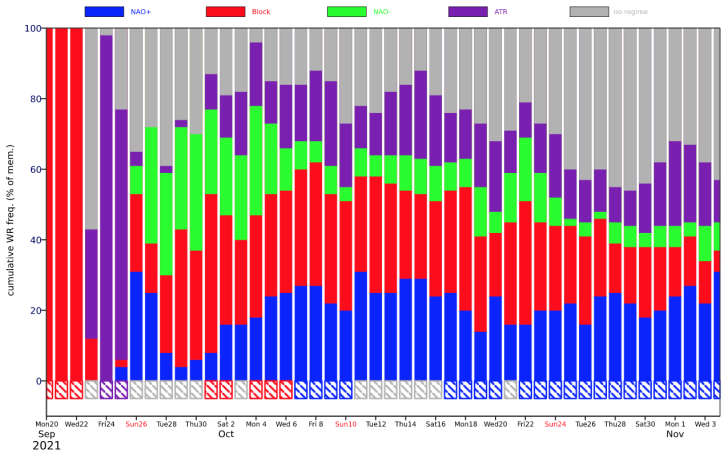


Figure 2.3: Visualization of the 45-days sub-seasonal forecast provided by ECMWF (Ferranti et al. 2011)

2.2 Data Pre-processing

The main issue of applying optimization techniques on geographical data refers to the high data dimensionality, especially whether vectorial and matrix computations are involved.
In the Machine Learning domain, this is often addressed as *curse of dimensionality*, i.e. the impossibility of efficiently learning the space where data samples are distributed, which becomes exponentially sparse as the dimensions of the feature space increase. For this reason, here comes the necessity to introduce some pre-processing techniques, able to create a suitable low-dimensional encoding, both minimizing the loss of information, and maximizing the performance of the optimization process.

2.2.1 Empirical Orthogonal Function (EOF) Analysis

Empirical Orthogonal Functions (EOFs) (Hannachi 2004) represents a standard reduction technique in geo-science-related experiments. Firstly adopted to reduce the feature space to few relevant variables (sharing a common concept with PCA, Section 2.2.2), it is mainly used to extract individual modes of variability (i.e. predominant patterns, often sharing common structures with the main regimes). There exist several variants of the algorithm, according if a spatial-time cross-correlation is kept into account; in this section, we refer to the standard algorithm, where the components are considered to be totally de-correlated (orthogonal). Although we know that the real physical components share some dependencies, we believe this is a suitable trade-off leading to a minor computational complexity.
Given the time-indexed geo-dataset, we can indicate it as:

$$\mathbf{X}_{t,y,x}, \quad t = 1, \dots, N$$

i.e., indexed by time t, latitude y, longitude x, respectively.
The EOF procedure requires a 2D-flattened version of this dataset, where the spatial indexes are converted from a grid to a vectorial format, leading to:

$$\mathbf{X}_{t,p}, \quad t = 1, \dots, N, \; p = (y_i, x_i), \; i = 1, \dots, P$$

Since data is provided in a regular grid format, there could be the necessity to introduce a weighting scheme prior to analyse data and fed it to EOF analysis. Indeed, the density given by the un-weighted grid, would be major poleward, than along the latitude. This attempt requires then the following mathematical passage:

$$\mathbf{X}^w = \mathbf{X} D_\theta, \quad D_\theta = \mathrm{diag}[\cos\theta_1, \dots, \cos\theta_P]$$

Finally, the Empirical Orthogonal Functions are computed solving the *Singular Value Decomposition* (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

factorizing the dataset in:

- $\mathbf{U} \in \mathbb{R}^{n \times n}$, whose columns are called *left singular vectors*, representing the eignevectors of the matrix $\mathbf{X}\mathbf{X}^T$, as well as the EOFs along the spatial dimension
- $\mathbf{\Sigma} \in \mathbb{R}^{n \times p}$: the matrix of the singular values
- $\mathbf{V} \in \mathbb{R}^{p \times p}$, whose columns are called *right singular vectors*, representing the eignevectors of the matrix $\mathbf{X}^T\mathbf{X}$, as well as the Principal Components (PCs) along the time dimension

By construction, EOFs are stationary (i.e. they do not evolve over time), only changing in the sign and value of their amplitude, thus resembling the state of the atmosphere.

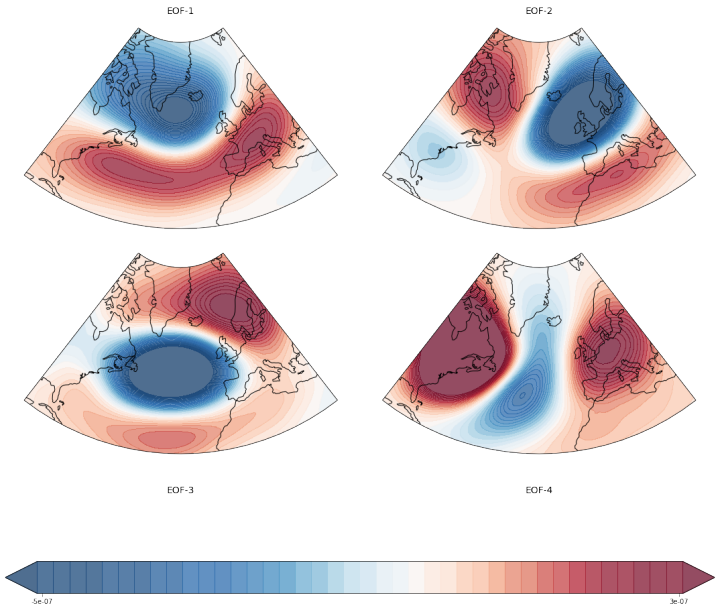


Figure 2.4: Visualization of the four main EOFs (Empirical Orthogonal Functions) in the North-Atlantic/European zone. It is possible to identify: a NAO+ resembling component (top left); an anti-Scandinavian Block SB- (top right); an anti-Atlantic Ridge AR- component (bottom left)

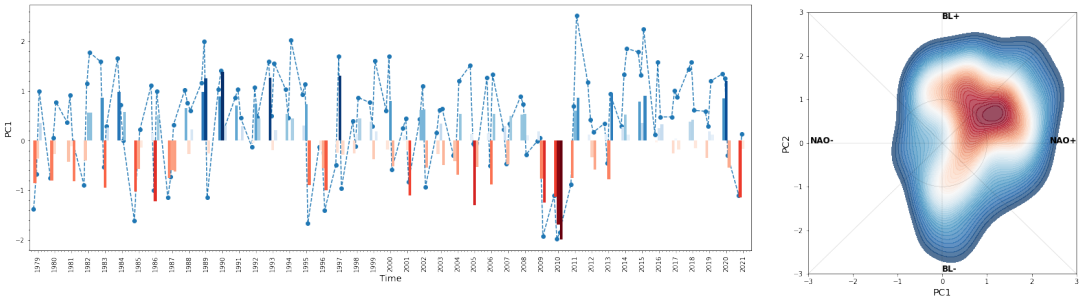


Figure 2.5: Bar Visualization (left) of the main Principal Component (PC-1) of the Time Series, in the North-Atlantic/European zone, grouped by the January-February-December months of the same solar year. PC is here standardized to have zero mean and unit variance. When the values are out of the range ± 1, they are generally associated to more extreme meteorological events. Associated to it, the NAO Index (Climate Prediction Center n.d.) evolution during winter months: it is evident the high correlation with PC-1. On the right, a density visualization of the first two PCs

2.2.2 Principal Component Analysis (PCA)

PCA represents a standard pre-processing technique for dimensionality reduction in Machine Learning problems. As for EOF analysis, the intent is to find a low-dimensional (latent) representation $\mathbf{Z}$ of data, through a projection matrix $\mathbf{W}$, such that:

$$\mathbf{Z} = \mathbf{XW}$$

by maximizing the variance (and thus, the explainability) of it. It is an optimization problem, described by the objective:

$$\underset{w:\ \|w\|<1}{\operatorname{argmax}} \|\mathbf{X}w\|_2^2$$

where we introduce a constraint on the projection weights w, in order to bound the problem and regularize the magnitude of them.
As a result, it translates into a closed-form analytical solution, coinciding with the eigen-decomposition problem. The projection matrix $\mathbf{W}$ is thus the matrix of eigen-vectors of $\mathbf{X}$. By picking the top-n eigenvectors, associated to the largest-magnituded eigenvalues, we can control the quantity of explained variance associated to the latent variables z.

2.2.3 Variational Autoencoder (VAE)

As clearly proved, the PCA approach performs a linear transformation of data into an encoded space. This feature represents both the strength and weakness of the method, able to computationally scale with large datasets, but often inferring a trivial mapping.
For this reason, we also provide a more advanced dimensionality reduction technique, leveraging the approximation power of neural networks.
Autoencoders (AEs) are a Deep Learning architecture, composed of two stages (Figure 2.6): an *encoder*, mapping original data into a low-dimensional representation, called *bottleneck* (in this case, we deal with *undercomplete* AEs); a *decoder*, learning to reconstruct the input samples from the latent space. The simplest architecture we can think of, is a 3-layers network. In the absence of non-linear activation functions, this architecture should converge to the PCA, being the bottleneck obtained through a matrix-to-matrix multiplication with the input.
In our study, we decide to tackle the problem in a more sophisticated way, directly inferring a mapping on the anomaly maps (Saenz et al. 2018), thus employing techniques from the computer vision domain (Prasad et al. 2020).
We switch to a probabilistic approach, employing a Variational Autoencoder (VAE) architecture. By definition, a VAE is an Autoencoder where the latent space summarizes a multi-variate distribution, learning its mean and covariance parameters

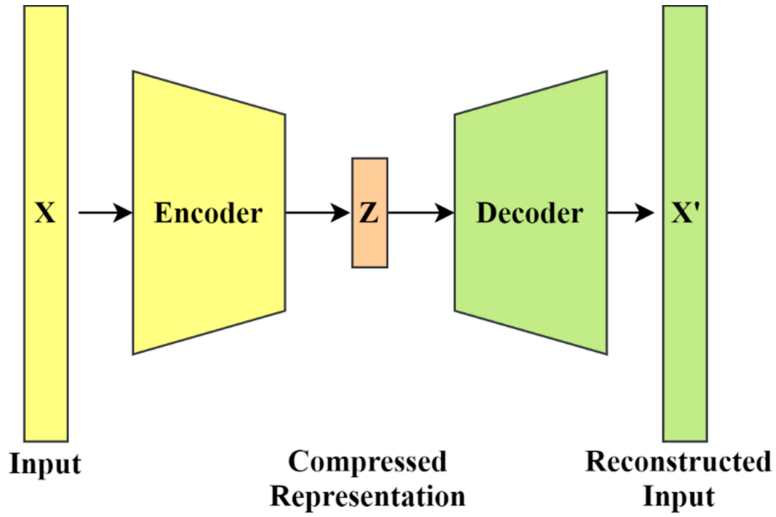


Figure 2.6: Visualization of an Autoencoder architecture(Elbattah et al. 2021)

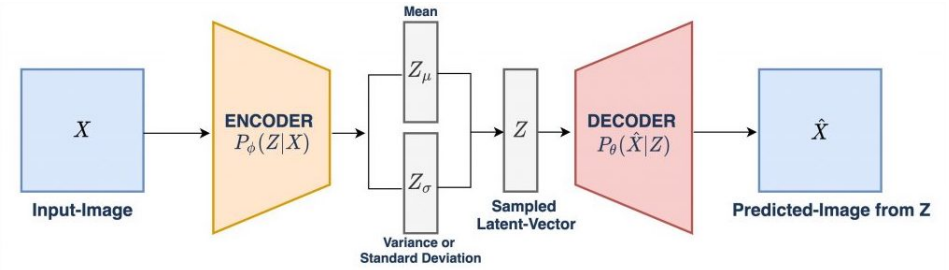


Figure 2.7: Visualization of a Variational Autoencoder (VAE) architecture(LearnOpenCV n.d.)

(Figure 2.7). The "variational" addressing is referred to the statistical approximation performed by the two networks, optimizing the so-called *variational objective*. Indeed, under a mathematical point of view, a VAE tries to find the distribution $p(z|x)$ which best explains data x starting from its low-dimensional representation z. Recalling the Bayes' rule we can fragment this conditional *posterior* by means of a *likelihood* $p(x|z)$ on the reconstructed data, a *prior* $p(z)$ on the latent space, and a *marginal likelihood* (often addressed as model evidence) $p(x)$. Therefore, according to the Bayes' theorem it follows that:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

Under a more practical point of view and with a slight change of notation, the decoder, parametrized by θ parameters, approximates the likelihood distribution $p_\phi(x|z)$, while the encoder (ϕ-parametrized) is in charge of approximating the posterior $p_\phi(z|x)$. Finally, the prior on the latent space is typically chosen to be as more general as possible, with a standard multivariate Gaussian representing the usual choice.
The framework presents an innermost issue, though. The marginal likelihood term $p(x)$ is indeed tough to be estimated analytically, leading to an intractable analytical solution. Hence, we resort to *variational inference*, introducing an approximating distribution $q(z)$, which we optimize to be as closer as possible to the true posterior distribution.
The objective statement then aims at reducing the gap between these two distributions, by minimizing their *Kullback-Leibler (KL-)Divergence*:

$$\underset{z,\ \phi,\ \theta}{\operatorname{argmin}}\ \mathbb{KL}(q_\phi(z)\|p_\phi(z|x))$$

By definition, we can write the KL-term as:

$$\underset{z,\ \phi,\ \theta}{\operatorname{argmin}}\ \mathbb{KL}(q_\phi(z)\|p_\phi(z|x)) =$$

$$= \underset{z,\ \phi,\ \theta}{\operatorname{argmin}} \int q_\phi(z) \log \frac{q_\phi(z)}{p_\phi(z|x)} = \underset{z,\ \phi,\ \theta}{\operatorname{argmin}}\ \mathbb{E}_{q_\phi(z)}\left[\log \frac{q_\phi(z)}{p_\phi(z|x)}\right]$$

Then, exploiting the logarithm properties:

$$\underset{z,\ \phi,\ \theta}{\operatorname{argmin}}\ \mathbb{E}_{q_\phi(z)}[\log q_\phi(z)] - \underbrace{\mathbb{E}_{q_\phi(z)}[\log p_\phi(z|x)]}_{\text{probabilistic term}}$$

Focusing on the probabilistic term, we can apply the Bayes' rule to simplify a bit the notation:

$$\mathbb{E}_{q_\phi(z)}[\log p_\phi(z|x)] = \int q_\phi(z) \log \left[\frac{p_\theta(x|z)p(z)}{p(x)}\right]$$

$$= \mathbb{E}_{q_\phi(z)} \log p_\theta(x|z) + \mathbb{E}_{q_\phi(z)} p(z) - \mathbb{E}_{q_\phi(z)} p(x)$$

Recalling the initial objective, where we intentionally omit the argmin notation for sake of brevity:

$$\mathbb{KL}(q_\phi(z)||p_\phi(z|x)) = \mathbb{E}_{q_\phi(z)}[\log q_\phi(z)] - (\mathbb{E}_{q_\phi(z)} \log p_\theta(x|z) + \mathbb{E}_{q_\phi(z)} p(z) - \mathbb{E}_{q_\phi(z)} p(x))$$

With a bit of re-arrangement, we can define the variational objective, also known as *Evidence Lower Bound (ELBO)*:

$$p(x) - \mathbb{KL}(q_\phi(z)||p_\phi(z|x)) = \mathbb{E}_{q_\phi(z)} \log p_\theta(x|z) - \underbrace{(\mathbb{E}_{q_\phi(z)}[\log q_\phi(z)] - \mathbb{E}_{q_\phi(z)} p(z))}_{=\mathbb{KL}(q_\phi(z)||p(z))}$$

Given the change of sign, this new objective is maximized, and it is equivalent to the maximization of the marginal likelihood metric, representing a lower bound on it. Indeed, by minimizing the KL-divergence between the approximating and true posteriors, we try to make the lower bound tighter. At the same time, the log-likelihood on the reconstructed data is maximized. The last term represents a *regularization* term, which explains the intended evolution of the optimization process: at the beginning the approximating distribution $q_\phi(z)$ is thought to be very close to the uninformative prior of the latent space $p(z)$. The scope is then to gradually shift the approximating distribution towards the more complex posterior. Under a more practical point of view, it is often hard to find the correct balance to reach the desired performances. Indeed, the risk is that over-parameterized model could lead to an excessive weight of the last term, often resulting in trivial or weakly-distinguished encodings. For this reason, over the years several variants of VAEs have been proposed, with the scope of introducing more dynamic objectives, with twofold purposes:

- leading to better reconstructions
- obtaining the so-called *disentangled* representations (Locatello et al. 2018; Locatello et al. 2019a; Locatello et al. 2019b; Locatello et al. 2020; Träuble et al. 2021; Dittadi et al. 2020; Mita et al. 2021)

While the former is often more related to image-reconstruction problems, where standard VAEs tend to produce blurry reconstructions, the latter has been at the center of many studies. We define disentagled representations as a suitably small set of latent features, uncorrelated between each other, and thought of being sufficiently informative for the generation process. As a standard example, characteristics like the geometry, shape, size, colour, orientation of an object in an image could be tought as disentangled features. Being able to retrieve them means finding a very compressed feature mapping, enhancing the overall data and model interpretabilities.

A drawback of this approach is that it often requires supervised information for at least few data samples, turning the problem into a supervised/weakly-supervised learning problem. In our study, this represents a big obstacle since we do not have at disposal any labeling. Also, there could not be the possibility to artificially re-create a pseudo-labeling, since each anomaly map could fall into multiple classes with different probabilities, making an hard-classification impossible.
However, we are able to find a trade-off, employing two well-noted VAE variants:

- β-VAE (Higgins et al. 2017): this variant introduces the hyper-parameter β in the variational objective, weighting the KL-term. It results then in:

  $$\underset{z,\ \phi,\ \theta}{\operatorname{argmin}}\ \log p_\theta(x|z) - \beta \cdot \mathbb{KL}(q_\phi(z)\|p(z))$$

  The adoption of a $\beta > 1$ should leverage the learning of disentangled representation, by forcing the latent variables in a more constrained/regularized space. As a trade-off we lose in reconstruction capabilities.
  In our experiments, although we encounter this effect of better decoupled latent features, the adoption of the β-VAE results in decreased performances of the subsequent clustering task in the pipeline. Indeed, the strong imposed regularization projects the points in a much more restricted region of the space, having as a side effect a distortion of the clustering validation measures based on the points dispersion

- σ-VAE (Rybkin et al. 2020), an architecture introducing the variance term of the likelihood as a trainable parameter to *calibrate* (in statistical sense) the decoder distribution. The variational objective is thus modified as follows:

  $$D \log \sigma + \frac{1}{\sigma^2} \log p_\theta(x|z) - \mathbb{KL}(q_\phi(z)\|p(z))$$

  where D is the input-data dimensionality.
  Conceptually, the decoder distribution (typically either a multivariate Gaussian or a Multinoulli) is set as:

  $$p_\theta(x|z) = \text{Dist}(\mu_\theta(z), \sigma^2_\theta(z))$$

  even though, in practice, imposing a common shared diagonal covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ works better.
  Furthermore, it allows to resort to the β-VAE objective, by setting the variance-term σ^2 constant and equal to $\frac{\beta}{2}$. In this way, there is no need to manually tune β.

We reach a noticeable improvement with respect to the PCA reduction, managing to encode the anomaly maps in a 5-dimensional space, ensuring major interpretability and cheaper computations in the subsequent modeling task.

To assess the improvement on the encoded space we rely on a twofold proof. First, the results we obtain for modeling weather regimes (Section 3.3) with the VAE algorithm are satisfactory and well in line with the literature, despite a much denser feature space. Thus, it suggests that the VAE algorithm is more efficient to separate the points belonging to different clusters. Secondly, we employ the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm to find a 2D representation of the high-dimensional manifold of the reduced space, visually proving the enhancements we reach with the VAE (Figure 2.8). Each point in the chart corresponds to a different anomaly map in our historical dataset, associated to a strong signal of a regime (i.e. where an hard-assignment to a singular regime would be a faithful approximation). The representation is clearly less sparse in terms of intra-cluster variance for the VAE reduction scheme than under the PCA projection. Further, we highlight also how in both the two projections the points associated to AR are typically mixed within the other regimes (more on this in the subsequent Sections).

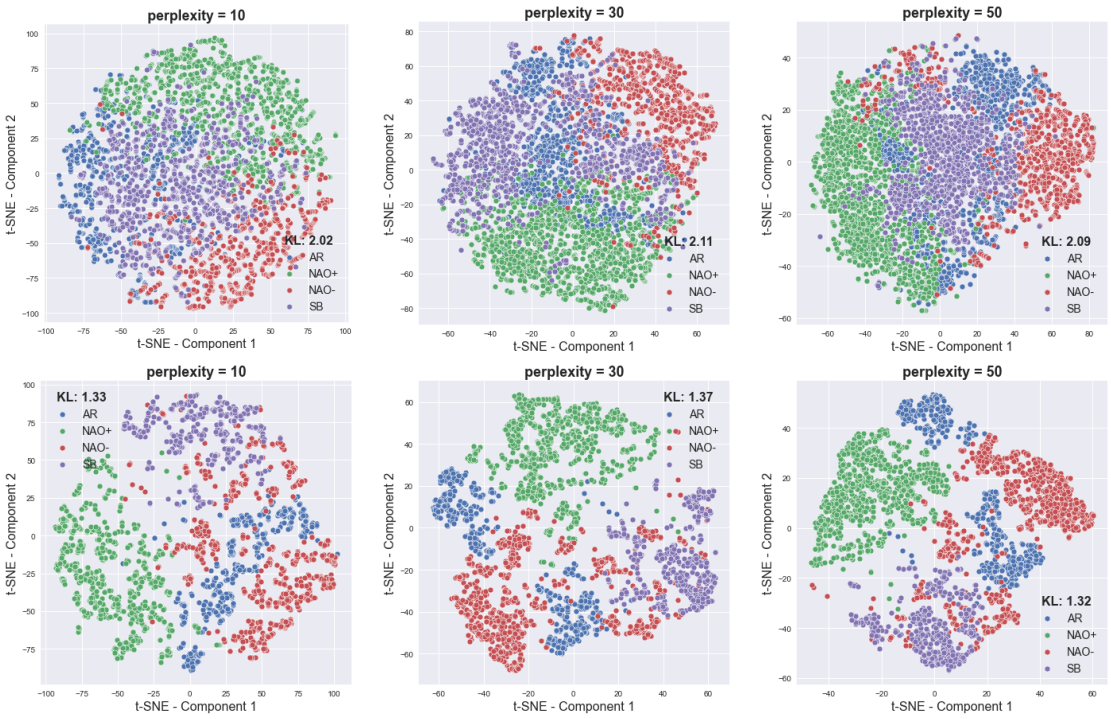


Figure 2.8: t-SNE applied on the PCA-reduced (top row) and VAE-reduced (bottom row) spaces at different levels of perplxities: the VAE-encoded features leads to a better separability of the clusters, as further proved by the better minimized KL-divergence objective

In order to guarantee a truthfulness of these considerations, such visualization is computed at different levels of *perplexity*, representing a parameter of the t-SNE algorithm determining the number of neighboring points used to estimate the projection of the manifold in the bi-dimensional space. To do so, the algorithm minimizes a KL-Divergence objective, whose value we use to further assess the better projection entailed by the VAE reduction.

Chapter 3

Models

We adopt a totally-unsupervised approach to the problem, given its particular nature. We initially inspect the standard K-Means algorithm (Section 3.1), in order to align with the previous studies on the topic (Grams et al. 2017; Falkena et al. 2020; Cassou et al. 2005; Cassou et al. 2011; van der Wiel et al. 2019a; van der Wiel et al. 2019b; Bloomfield et al. 2016). In literature the experiments only focus on a frequentist approach to the clustering of the regimes, intended to give an high level overview of the most prominent configurations. However, we think that a daily-level hard-assignment to a particular regime represents an excessively simplistic assumption, although driven by ease of implementation.
Further, we see that across the studies (Grams et al. 2017; Falkena et al. 2020; Cassou et al. 2011; van der Wiel et al. 2019a) a discrepancy exists about the correct number of clusters to be considered, despite the four, well-established, theoretical configurations are often enough to accurately track the atmospheric dynamics. This drawback could also be addressed to a wrong choice of the algorithm, that necessarily has to take into account higher numbers of clusters, to track intermediate evolutions of the regimes, thus performing a more faithful assignment. For this reason, after some experiments on the framework, we decide to switch to a fully-Bayesian approach, adopting Mixture Models (Section 3.2) (Vrac et al. 2007; Smyth et al. 1999; Stan et al. 2017).
The major complexity we introduce is justified by the following advantages:

1. Major statistical interpretability of the system and its dynamics

2. A direct choice of the number of the clusters: we can indeed either opt for the standard solution of 4 clusters, where the intermediate transitions between mixed regimes are caught by the assignments of more uniform probabilities to multiple clusters, or we can rely on ad-hoc validation measures taken directly from the probabilistic setting

3. A real descriptive probabilistic weighting scheme to the occurrence of the regimes, not based on a simple frequency

4. Linked to the previous point, we can infer a more truthful and calibrated quantification of the subsequent tasks in the pipeline, considering the effects deriving from multiple configurations, weighted by the related probabilities

3.1 K-Means

K-Means is a standard clustering algorithm, exploiting a distance metric to cluster points in K pre-defined clusters. Its ease of implementation makes it a standard choice for most of the clustering tasks, and the inference only requires to store the positions of the centroids, i.e. the clusters' centers as obtained after the training. The iterative algorithm (*Expectation-Maximization (EM)* algorithm) can be summarized as follows:

Algorithm 1 K-Means

Initialize centroids $\mu_k,\ k = 1, \dots, K$
loop
  loop over each point $x_n,\ n = 1, \dots, N$
    Evaluate cluster assignment as: ▷ **E**-step

$$z_{nk} = \underset{k}{\operatorname{argmin}}\ (x_n - \mu_k)^T (x_n - \mu_k)$$

    Recompute centroids: ▷ **M**-step

$$\mu_k = \frac{\sum_{n=1}^{N} z_{nk} x_n}{\sum_{n=1}^{N} z_{nk}},\ k = 1, \dots, K$$

until convergence

As most of the clustering algorithms, initialization and convergence criteria are two critical factors in determining the overall performances. While the latter is typically bypassed by setting an appropriate number of iterations, the former requires several different runs of the algorithm, possibly converging to as many different solutions.
Some other limitations are then referred to:

- the adoption of the Euclidean distance metric, treating all the dimensions of the feature space evenly. The relative representativeness and importance of a feature is completely lost

- non-linear separable clusters, even though clearly recognizable, are often wrongly recognized

3.2 Mixture Models

Mixture models are generative models focusing on how data could have created. Each point x_n is supposed to be generated from one of the K distributions, that can be easily sampled to generate new points.
Given the set of the datapoints $\mathbf{X}$, we define as Δ the set of all the parameters of the distributions. The model is a reverse-engineering process, aiming to learn Δ, and thus the originating distribution associated to each point.
The objective is to maximize the *mixture model likelihood*. Let the k-th distribution have probability density function:

$$p(x_n|z_{nk} = 1, \Delta_k)$$

The marginal likelihood associated to data $\mathbf{X}$ is $p(\mathbf{X}|\Delta)$.
We leverage the i.i.d. (independent and identically distributed) assumption over the data-points $x_n \in \mathbf{X},\ n = 1, \dots N$, therefore factorizing:

$$p(X|\Delta) = \prod_{n=1}^{N} p(x_n|\Delta)$$

Then, we un-marginalize k:

$$p(\mathbf{X}|\Delta) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(x_n, z_{nk} = 1|\Delta) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(x_n|z_{nk} = 1, \Delta_k)p(z_{nk} = 1|\Delta)$$

Hence, since we want to find Δ maximizing it:

$$\underset{\Delta}{\operatorname{argmax}} \prod_{n=1}^{N} \sum_{k=1}^{K} p(x_n|z_{nk} = 1, \Delta_k)p(z_{nk} = 1|\Delta)$$

We switch to log-probabilities to simplify the optimization:

$$\underset{\Delta}{\operatorname{argmax}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(x_n|z_{nk} = 1, \Delta_k)p(z_{nk} = 1|\Delta)$$

However, applying a logarithm of a sum could be tricky to be optimized, then we resort to the *Jensen's Inequality*:

$$\log \mathbb{E}_{p(x)}[f(x)] \geq \mathbb{E}_{p(x)}[\log f(x)]$$

So, given our likelihood $\mathcal{L} = \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(x_n|z_{nk}=1,\Delta_k)p(z_{nk}=1|\Delta)$, we add a (arbitrary looking) distribution $q(z_{nk}=1)$, subject to $\sum_k q(z_{nk}=1)=1$:

$$\mathcal{L} = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \frac{q(z_{nk}=1)}{q(z_{nk}=1)} p(x_n|z_{nk}=1,\Delta_k)p(z_{nk}=1|\Delta)$$

We now have an expectation:

$$\mathcal{L} = \sum_{n=1}^{N} \log \mathbb{E}_{q(z_{nk}=1)} \left[\frac{1}{q(z_{nk}=1)} p(x_n|z_{nk}=1,\ \Delta_k)p(z_{nk}=1|\Delta) \right]$$

telling us how on average the parameters of $q(z_{nk})$ are able to model the likelihood of data, i.e. how on average the produced assignments explain the data disposition in the space. If we look at $q(z_{nk})$ as a posterior distribution (i.e. $q(z_{nk}|x_n)$), the equation could be interpreted as:

$$\mathbb{E}_{\text{posterior}}\left[\frac{\text{likelihood}\cdot\text{prior}}{\text{posterior}}\right] = \mathbb{E}_{\text{posterior}}[\text{marginal likelihood}]$$

So, applying the Jensen's Inequality:

$$\mathcal{L} \geq \sum_{n=1}^{N} \mathbb{E}_{q(z_{nk}=1)} \left[\log \frac{1}{q(z_{nk}=1)} p(x_n|z_{nk}=1,\ \Delta_k)p(z_{nk}=1|\Delta) \right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1) \log \left\{ \frac{1}{q(z_{nk}=1)} p(x_n|z_{nk}=1,\ \Delta_k)p(z_{nk}=1|\Delta) \right\}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1) \log p(z_{nk}=1|\Delta) +$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1) \log p(x_n|z_{nk}=1,\ \Delta_k) -$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1) \log q(z_{nk}=1)$$

We then define $q_{nk} = q(z_{nk}=1),\ \pi_k = p(z_{nk}=1|\Delta)$ (both just scalar quantities):

$$\mathcal{L} \geq \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \log p(x_n|z_{nk}=1,\ \Delta_k) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk} \log q_{nk}$$

and we differentiate this lower bound with respect to $q_{nk},\ \pi_k,\ \Delta_k$, setting it to zero and obtaining an iterative update.

The updates for $\Delta_k,\ \pi_k$ will depend on q_{nk}, therefore the latter will be firstly updated.
This is another form of the Expectation-Maximization algorithm, derived in a different way. Again, the algorithm is very sensitive to initialization, however it is guaranteed to converge to a local maximum of the lower bound.
The overall algorithm can be summarized as:

Algorithm 2 Mixture Model

Guess $\mu_k,\ \sigma_k^2,\ \pi_k\ k = 1, \ldots, K$
loop
  Compute $q_{nk},\ n = 1, \ldots, N,\ k = 1, \ldots, K$
  Update $\mu_k,\ \sigma_k^2,\ k = 1, \ldots, K$
until convergence

The clustering assignment step is governed by the probabilities q_{nk} of a point x_n coming from the distribution k:

$$q_{nk} = p(z_{nk} = 1 | x_n, \mathbf{X})$$

Each point will be then characterized by a set of K probabilities.

3.2.1 Gaussian Mixture Model (GMM)

We assume the component distributions are Gaussians with diagonal covariance, $p(x_n | z_{nk} = 1,\ \mu_k,\ \sigma_k^2) = \mathcal{N}(\mu,\ \sigma^2 I)$.
When updating π_k we are subjected to the constraint $\sum_k \pi_k = 1$, so we need to add a Lagrangian to the update term:

$$\sum_{n,k} q_{nk} \log(\pi_k) - \lambda \Big(\sum_k \pi_k - 1 \Big)$$

Differentiating it and setting it to zero:

$$\frac{\partial}{\partial \pi_k} = \frac{1}{\pi_k} \sum_n q_{nk} - \lambda = 0 \longrightarrow \sum_n q_{nk} = \lambda \pi_k$$

Summing both sides over k, we find

$$\lambda = \sum_{n,k} q_{nk}$$

Therefore, substituting:

$$\pi_k = \frac{\sum_n q_{nk}}{\sum_{n,j} q_{nj}} = \frac{1}{N} \sum_n q_{nk}$$

Update for q_{nk}: now the whole bound is relevant. Adding the Lagrange term $-\lambda\Big(\sum_k q_{nk} - 1\Big)$ and differentiating:

$$\frac{\partial}{\partial q_{nk}} = \log \pi_k + \log p(x_n|z_{nk} = 1,\ \Delta_k) - (\log q_{nk} + 1) - \lambda$$

Re-arranging ($\lambda' = f(\lambda)$):

$$\pi_k p(x_n|z_{nk} = 1,\ \Delta_k) = \lambda' q_{nk}$$

Summing over k to find λ' and re-arranging:

$$q_{nk} = \frac{\pi_k p(x_n|z_{nk} = 1,\ \Delta_k)}{\sum_{j=1}^{K} \pi_j p(x_n|z_{nj} = 1,\ \Delta_j)}$$

where evaluating $p(x_n|z_{nk} = 1,\ \Delta_k)$ requires to plug x_n inside a multi-variate Gaussian parametrized by μ_k, σ_k^2.
Updates for $\mu_k,\ \sigma_k^2$: these are easier (they are not subjected to any constraint). Differentiating the following and setting to zero:

$$\sum_{n,k} q_{nk} \log \left\{ \frac{1}{(2\pi\sigma_k^2)^{D/2}} \exp\left(-\frac{1}{2\sigma_k^2}(x_n - \mu_k)^T(x_n - \mu_k) \right) \right\}$$

$$\mu_k = \frac{\sum_n q_{nk} x_n}{\sum_n q_{nk}}, \quad \sigma_k^2 = \frac{\sum_n q_{nk}(x_n - \mu_k)^T(x_n - \mu_k)}{D \sum_n q_{nk}}$$

3.2.2 Dirichlet Process Mixture Model (DPMM)

Dirichlet Process Mixture Models (DPMM) (Blei et al. 2006) is another probabilistic model, leveraging the framework up to an infinite number of mixtures. It is a *non-parametric* model, meaning that the number of parameters can grow with higher amounts of data. This is especially useful when the number of clusters is not fixed, or there is the necessity to extrapolate it from data. The algorithm is Bayesian, setting a probabilistic distribution over the number of clusters.
To introduce it, we firstly define a *Dirichlet Process* $\mathcal{DP}(\alpha, G)$, as a distribution over distributions, where:

- α is the concentration parameter: the smaller it is, the fewer points result with high probabilities
- G is the base distribution we want to shape through the process itself

Under a practical point of view, we can refer to a DP as a black-box having as input the distribution G, and as output a distribution G', whose similarity to G is determined by α.

$$G \rightarrow \boxed{\mathcal{DP}_\alpha} \rightarrow G'$$

Instead, formally, the underlying statistical process estimates the break-sticking weights C_k as a recursive sampling from a Beta distribution:

$$\pi_k = V_k \prod_{j=1}^{k-1}(1 - V_k), \quad \text{where} \quad V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha)$$

Given these weights, we can therefore approximate the original distribution G by drawing some elements from it:

$$\mu_1, \mu_2, \cdots \sim_{\text{iid}} G$$

and merging them into a complete distribution:

$$G = \sum_{k \in N} \pi_k \delta_{\mu_k}$$

where δ_{μ_k} is a Dirac's delta (i.e. an indicator function) associating each drawn element μ_k with its corresponding weight π_k.
Some important properties associated to a Dirichlet Process are then:

- Invariance of the expected value of the distribution G:

$$\mathbb{E}_{\mathcal{DP}(\alpha,G)}[x] = \mathbb{E}_G[x]$$

- Convergence at the original distribution G at the bounds, i.e. as $\alpha \rightarrow \infty$, $\mathcal{DP}(\alpha, G) = G$
- Composition of DPs:

$$H \sim \mathcal{DP}(\alpha, G)$$
$$J \sim \mathcal{DP}(\alpha, H)$$

Finally, we can define DPs as Mixture Models. We consider a Gaussian as a base distribution G, and we draw some elements, that will become the cluster centers. Since we can draw an unbounded number of elements, the number of clusters is possibly unbounded. The model will be then in charge to weigh each of the infinite clusters, eventually excluding the majority of them and adapting to the data $\mathbf{X}$.
To explain how inference works, we resort to a typical metaphor called the "The Chinese Restaurant". According to this, given a certain number of tables, to generate an observation we need to sit at a table, with a probability proportional

to the number of people sitting at that table. Each table will be then identified by a value (i.e. a statistical moment, which is the cluster center/mean vector in our case).
As we can easily see, this concept is funded on a Maximum Likelihood approach. However, in order to stick with a Bayesian approach, the framework introduces the concept a posterior (itself, a Dirichlet Process) by specifiying a prior of the same nature, allowing to reach the conjugacy.
A new observation will be then assigned to a new cluster proportionally to α, however being firstly balanced against the probability of an observation given a cluster. To do this, we typically consider the data-points as Normally distributed and belonging to K different clusters. At the beginning we assume there is no leakage of information about the separability of the clusters, and that we have unknown prior probabilities π_k of a particular data point belonging to the cluster k. Instead of imagining that each data point is firstly assigned a cluster and then drawn from the distribution associated to that cluster, we now think of each observation being associated with parameter μ_k drawn from some discrete distribution G. That is, we are now treating the μ_k as being drawn from the random distribution G, and our prior information is incorporated into the model by the distribution over distributions G. Extending this reasoning to an infinite number of clusters means selecting a random prior distribution:

$$G(\mu_k) = \sum_{k \in N} \pi_k \delta_{\mu_k}$$

More practically, we take a random guess initially, providing a mean for each of the K clusters. We incrementally change all of our assignments, unassigning and "questioning" each observation (in what is known as Gibbs Sampling mechanism). We do so by computing the following probability:

$$p(z_i | z_1, \ldots, z_{i-1}, z_{i+1}, z_N, \mathbf{X}, \alpha, G) = p(z_i | \{z_j\}_{j=1,\ldots,N \cap j \neq i}, \mathbf{X}, \alpha)$$

representing the probability of a clustering assignment z_i of a point x_i, given all the clustering assignments of the neighbouring points, and the parameters of the DPMM. We keep doing this for each data point, and it is proved to converge to the true posterior distribution.
Developing a bit the math, and defining as μ_k the means of the generating distributions, we can write the former probability (by means of the chain rule) as:

$$p(z_i = k | \{z_j\}_{j=1,\ldots,N \cap j \neq i}, \mathbf{X}, \{\mu_q\}_{q=1,\ldots,K}, , \alpha)$$

$$= p(z_i = k | \{z_j\}_{j=1,\ldots,N \cap j \neq i}, x_i, \{x_j\}_{j=1,\ldots,N \cap j \neq i}, \mu_k, \alpha)$$

$$= p(z_i = k | \{z_j\}_{j=1,\ldots,N \cap j \neq i}, \alpha) p(x_i | \mu_k, \{x_j\}_{j=1,\ldots,N \cap j \neq i})$$

$$= \begin{cases} \left(\frac{N_k}{N+\alpha}\right) \int_\mu p(x_i|\mu)p(\mu|G,\mathbf{X}) & \text{existing clusters} \\ \alpha \int_\mu p(x_i|\mu)p(\mu|G) & \text{new cluster} \end{cases}$$

where:

- we assume that all the distributions in the integral operator are Normal-like
- we see that the assignment weights $\left(\frac{N_k}{N+\alpha}\right)$ is explicitly proportional to the number N_k of data points inside a particular cluster, over the total number of points N

It turns out that everything can be simplified up to:

$$= \begin{cases} \left(\frac{N_k}{N+\alpha}\right)\mathcal{N}\left(x, \frac{N}{N+1}\mathbf{X}, \Sigma\right) & \text{existing clusters} \\ \alpha\mathcal{N}(x, 0, \mathbf{I}) & \text{new cluster} \end{cases}$$

when assuming zero-normal distribution with identity covariance matrix.
The Gibbs Sampling mechanism thus represents an approximation for the estimation of the infinite-dimensional posterior distribution $p(\pi, \mu|\mathbf{X})$ (i.e., sticking with an infinite number of clusters, we would need to associate to each data point an infinite-dimensional set of prior probabilities, and cluster means as well).
Another solution is to resort to Variational Inference, as we do in the case of GMMs. The underlying mathematical process is the same: an approximating distribution q for the posterior is introduced, and a lower bound on the marginal likelihood is optimized.
Despite the performant nature of the Gibbs Samplings, in our study we adopt this second approach, being less sensitive to initialization and typically faster to converge.

3.3 Results

For the training session of the models, we resort to a standard 5-Fold Cross-Validation process, performing hyper-parameter tuning through a grid search mechanism. This procedure is particularly beneficial to infer the correct number of clusters that best approximates the data distribution. Indeed, the sole maximization of certain metrics would often result in the trivial solution of adopting as many clusters as possible, thus overfitting data.
We guarantee the consistency of each tested combination by performing five (in the case of probabilistic models) to ten re-initializations of the initial configuration, given the relevant importance assumed by the starting phase.

We collect a set of different metrics to evaluate the quality of the clustering. Unfortunately, since we act in a completely unsupervised context, where adopting any sort of ground truth would be biased by the underlying modeling assumption. For this reason, the Météo-France datasets is adopted as a monitor of the performances of the clustering algorithms, however without being involved in the model selection process. Hence, we can only recur to the so-called *internal* clustering scores (in contrast to *external* metrics, leveraging some knowledge on some labelled data). These measures embed concepts of intra and inter-cluster dispersions, in order to evaluate the goodness of the assignment of each sample.

- *Inertia* or *Sum of Squared Errors (SSE)*: it is the typical measure when evaluating the clustering configurations. The measure indicates a sum of all the squared residuals between each point, and the cluster's centroid to which the point is assigned:

$$I = \sum_{k=1}^{K} \sum_{n=1:\ z_{nk}=1}^{N} (x_n - \bar{x}_k)^2$$

- *Silhouette Score*: a measure ranging in the interval $[-1, 1]$, evaluating the position of each sample with respect to the cluster which is assigned to, and the closest cluster. For each point it is evaluated:
  - *Mean dissimilarity* between a point x_i and the nearest cluster C_k:

$$b(x_i) = \min_{k \neq i} \sum_{j \in C_k} d(i, j)$$

  - *Mean intra-cluster distance* between a point x_i and all the other points belonging to its cluster:

$$a(x_i) = \frac{1}{|C_i| - 1} \sum_{j \neq i} d(i, j)$$

  The silhouette score for a point x_i is then:

$$s(x_i) = \begin{cases} 0 & \text{if } |C_i| = 1 \\ \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} & \text{otherwise} \end{cases}$$

  The overall silhouette score is then:

$$\text{Silhouette} = \max_k \bar{s}_k$$

  i.e., the maximum between the mean scores per cluster

- *Calinski-Harabasz (CH) Index*: a measure keeping into account the intra and inter-clusters variations

$$CH(k) = \frac{B(k)}{W(k)} \cdot \frac{n-k}{k-1}$$

  where:

  - $B(k)$ is the between-cluster variation, having $k-1$ degrees of freedom
  - $W(k)$ is the within-cluster variation, having $n-k$ degrees of freedom

- *Bayesian Information Criterion (BIC)* (Falkena et al. 2020): defined under the maximum likelihood estimation framework, it is evaluated as:

$$BIC = k \log N - 2 \log \hat{L}, \quad \hat{L} = p(x|\hat{\theta}_{MLE})$$

- *Evidence Lower Bound (ELBO)*: equivalent to the pre-defined variational objective

Table 3.1 summarizes the different metrics.

Model	**Inertia** ↓	**Silhouette score** ↑	**Calinski Harabasz Index** ↑	**Bayesian Information Criterion** ↑	**ELBO** ↑
K-Means	✓	✓	✓	✓	
GMM	✓	✓	✓	✓	✓
DPMM	✓	✓	✓	✓	✓

Table 3.1: Evaluation Metrics for each model. ↑/↓ indicates that the score has to be maximized/minimized

3.3.1 Theoretical Results

The process of Cross-Validation allows to select the best performing models. Table 3.2 reports the results for each algorithm, under each dimensionality reduction technique, including which score resulted the optimized objective.
At this point, we encounter the first critical points of this modeling step:

- The two dimensionality reduction techniques cannot be compared just looking at the absolute values of the score. Being dependent by concept of dispersion in the space, it is clear that a reduced space mapped on a less sparse region would result in higher scores. Due to a minor dimensionality of the feature space, the VAE-reduced space is more dense than the PCA hyper-space, thus leading to lower absolute scores

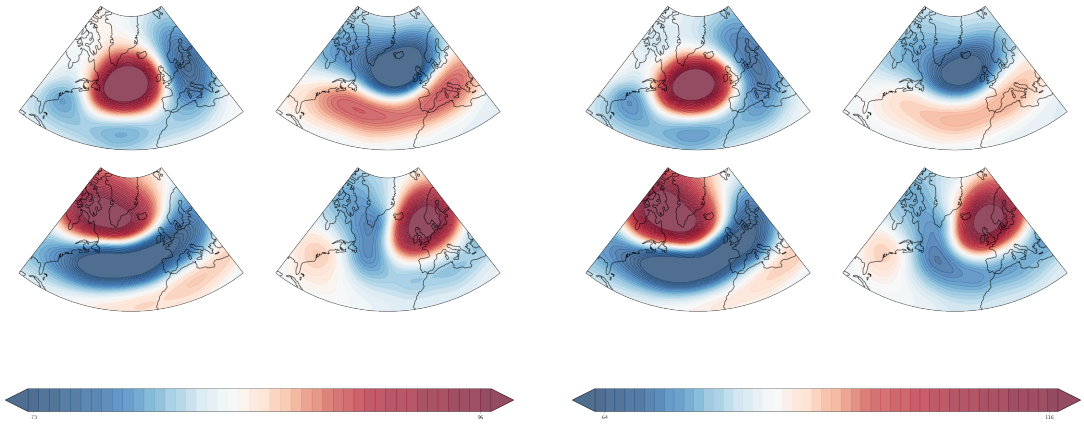
- The scores are a good metric to select the best model across several configurations of the same algorithm. However, as proven consequently, they fail in objectively determining which model, among the different assumptions, should be adopted as the reference candidate. While the two Mixture models are a closer assumption, the same does not happen when comparing K-Means with the probabilistic frameworks. With the exception of the ELBO, these metrics require a "deterministic" setting, like an hard thresholding of the assignments, thereby partially mitigating the generalization power of the probabilistic models

	Model	**Objective**	**Inertia** ↓	**Silhouette Score** ↑	**Calinski Harabasz Index** ↑	**Bayesian Information Criterion** ↑
PCA	**K-Means**	CH	5.27×10^{11}	0.113	93.4	-1.2×10^{-5}
	GMM	Silhouette	5.71×10^{11}	0.076	67.4	-1.2×10^{-5}
	DPMM	BIC	5.77×10^{11}	0.060	62.8	-1.2×10^{-5}
VAE	**K-Means**	CH	4.66×10^{7}	0.268	233.6	-2.46×10^{-4}
	GMM	CH	5.24×10^{7}	0.199	180.8	-2.49×10^{-4}
	DPMM	Silhouette	5.25×10^{7}	0.199	180	-2.49×10^{-4}

Table 3.2: Best performing models, under the two dimensionality reductions. ↑/↓ indicates that the score has to be maximized/minimized. The column "Objective" indicates which score was optimized to in the Cross-Validation process of the model

In Table 3.2 are reported only 4-clusters models, as we believe they are the correct trade-off to describe the regimes' dynamics, while not losing in generalization power. However, some 7-clusters configurations (Figure 3.2) under the K-Means choice are obtained from the model selection experiments, supporting the results of some previous studies (Grams et al. 2017). Figures 3.1 report the centroids configurations obtained with K-Means with the PCA (Figure 3.1a) and VAE (3.1b) reduction schemes. The inferred clusters' centers are very close to the theoretical ones, on the one hand confirming how globally K-Means could be adequate in describing the average dynamics. However, as Table 3.3 reports, a disagreement exists between the computed frequencies of the regimes in literature (Cassou et al. 2011; van der Wiel et al. 2019a; Luo et al. 2012; Meteo-France n.d.) and our experiments on K-Means. In particular, the major differences are highlighted into the configuration modeling on the VAE mapping.

Focusing instead on the predicted centroids of the Mixture Models (Figure 3.3) we immediately recognize some differences, with respect to K-Means, in the minor configurations, namely: AR and SB. This is a critical point we encounter,



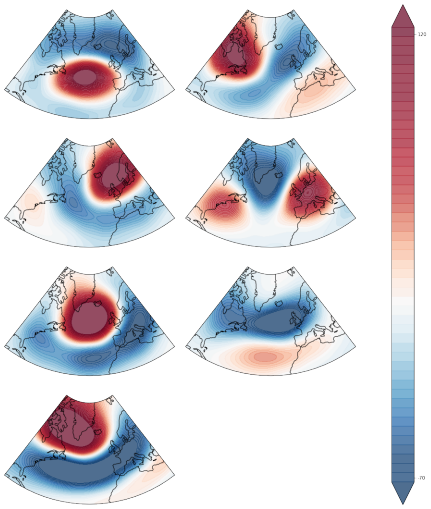
(a) Visualization of the centroids after clustering with K-Means on PCA-reduced dataset

(b) Visualization of the centroids after clustering with K-Means on VAE-reduced dataset

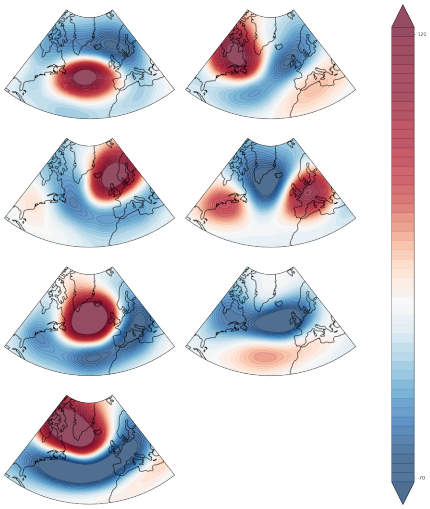
Figure 3.1: K-Means with 4 clusters. AR (top left), NAO+ (top right), NAO- (bottom left), SB (bottom right)

	Model	AR	NAO+	NAO-	SB
PCA	**K-Means**	20.6%	33.2%	19.9%	26.3%
VAE	**K-Means**	17.5%	47.4%	13.9%	21.2%
(Cassou et al. 2011)		23.3%	29.9%	22.4%	24.5%
(van der Wiel et al. 2019a)		20%	33%	20%	28%
(Luo et al. 2012)		21.8%	31.9%	18.1%	28.2%
(Meteo-France n.d.)		20.4%	31.8%	18.4%	29.4%

Table 3.3: Frequencies of the regimes as predicted by (Cassou et al. 2011; van der Wiel et al. 2019a; Luo et al. 2012; Meteo-France n.d.) vs. Frequencies as predicted by K-Means under PCA and VAE reduction schemes. We notice how K-Means is highly sensitive to the inferred mapping of VAE



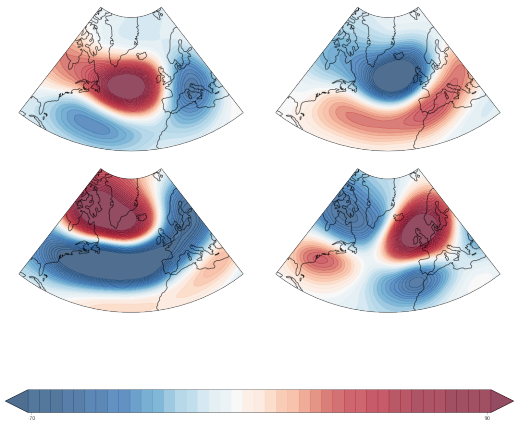
(a) Visualization of the centroids after clustering with K-Means on PCA-reduced dataset. The fine-tuning in this case emitted 7 clusters as the best configuration. Other than the four main standard configurations, it is possible to identify: SB- (top right), AR - (second row, on the right). The AR configuration can instead be conceived as average case between the top-left and third row, on the left plots



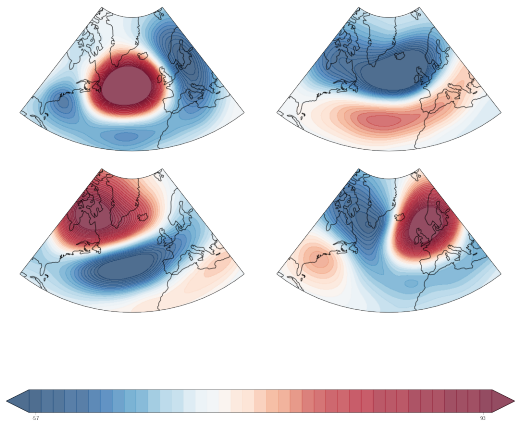
(b) Visualization of the centroids after clustering with K-Means on VAE-reduced dataset. The fine-tuning in this case emitted 7 clusters as the best configuration. Other than the four main standard configurations, it is possible to identify: SB- (top right), AR - (second row, on the right). The AR configuration can instead be conceived as average case between the top-left and third row, on the left plots

Figure 3.2: K-Means with 7 clusters

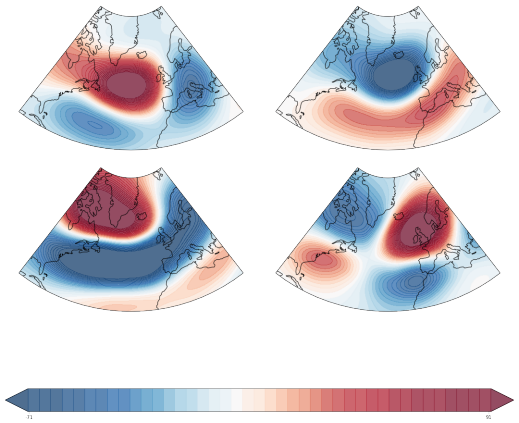
which is still more highlighted in the next section, when compared to Météo-France predictions, suggesting a small bias of these modeling choices.



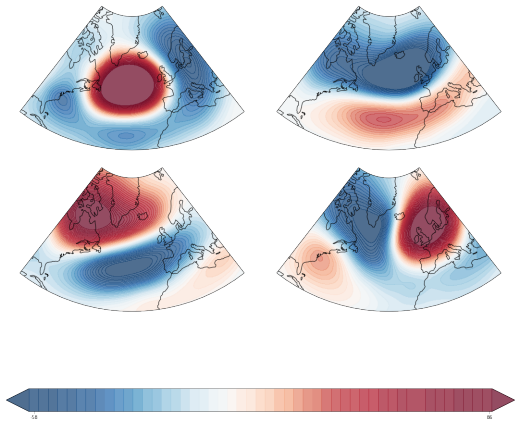
(a) Visualization of the centroids after clustering with GMM on PCA-reduced dataset



(b) Visualization of the centroids after clustering with GMM on VAE-reduced dataset



(c) Visualization of the centroids after clustering with DPMM on PCA-reduced dataset



(d) Visualization of the centroids after clustering with DPMM on VAE-reduced dataset

Figure 3.3: Mixture Models with 4 clusters. AR (top left), NAO+ (top right), NAO- (bottom left), SB (bottom right)

Again, we recur to an a-posteriori validation with the expected frequencies, as reported by the literature. Table 3.4 confirms that Mixture Models tend to benefit more from the robust reduction obtained through the σ-VAE, also driven by the

major ease of fitting multi-variate distributions on a minor dimensionality of the space.

	Model	AR	NAO+	NAO-	SB
PCA	**GMM**	27.9%	38.5%	15.5%	18.1%
	DPMM	27.1%	39.9%	14.8%	18.3%
VAE	**GMM**	17.7%	31.6%	26.9%	23.8%
	DPMM	24.9%	32.3%	18.1%	24.8%
(Cassou et al. 2011)		23.3%	29.9%	22.4%	24.5%
(van der Wiel et al. 2019a)		20%	33%	20%	28%
(Luo et al. 2012)		21.8%	31.9%	18.1%	28.2%
(Meteo-France n.d.)		20.4%	31.8%	18.4%	29.4%

Table 3.4: Frequencies of the regimes as predicted by (Cassou et al. 2011; van der Wiel et al. 2019a; Luo et al. 2012; Meteo-France n.d.) vs. Frequencies as predicted by Mixture Models under PCA and VAE reduction schemes. We notice how under PCA statistics tend to be more misaligned

The soft-assignments allow us to infer the intermediate transitions, thus excluding the necessity of adopting models with higher numbers of clusters (Cassou et al. 2011; Grams et al. 2017). Table 3.5 reports the frequency where two regimes are mixed together with similar probabilities (i.e. when they account for at least 70% of the total assigned probabilities). The results are quite realistic, since, empirically, positive NAO and SB are likely to merge as it will be also more deeply outlined in Sections 3.3.4 and 3.3.3.

	AR	NAO+	NAO-	SB
AR	-	2.4%	2.4%	2.6%
NAO+	2.4%	-	3%	**6.3%**
NAO-	2.4%	3%	-	3.3%
SB	2.6%	**6.3%**	3.3%	-

Table 3.5: Frequencies (%) of the intermediate transitions between pairs of regimes, as predicted by our Mixture Models.

Finally, as for K-Means, some configurations of Dirichlet Process Mixture Model allocate mass on 7-clusters scenarios (Figure 3.4), resulting a sub-optimal solution under the illustrated scores. Differently than K-Means though, the 7 regimes results to be only partially descriptive of the 7 theorized regimes of (Grams et al. 2017), clearly representing noisy intermediate transitions. While this experiment is not

further studied, we believe it could be of future interest analyzing these 7 clusters, possibly explaining at a finer-grained level some perspectives on weather regimes.

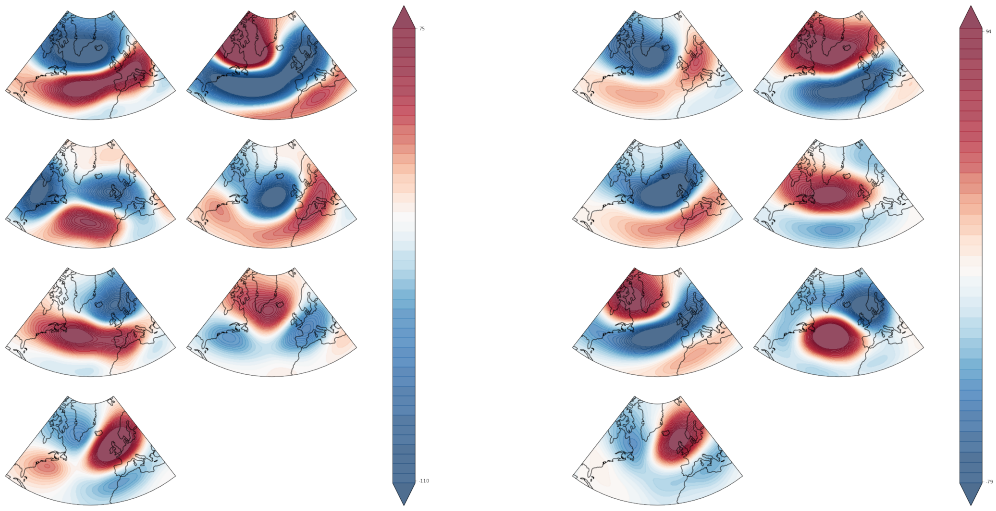


Figure 3.4: Variational Dirichlet Process Mixture Model with 7 clusters

3.3.2 Comparison with Météo-France

During our empirical validation of random sampled predictions we notice some discrepancies referred to K-Means predictions. However such analysis would not be feasible when enlarged to the entire dataset. Hence, we resort to a thorough comparison with Météo-France historical predictions, allowing us to circumscribe our modeling choices for the successive tasks. The models we select are K-Means under PCA reduction, and the Mixture Models under the VAE pre-processing. As Figure 3.5 reports, K-Means shows consistent performances, since it is aligned with the same modeling assumption of Météo-France. Mixture Models tend to be inline with the provider's data, however evidencing some insights:

- North Atlantic Oscillations (NAOs) configurations present an higher True Positive Rate (TPR). This is coherent if we think that they typically correspond to the major source of variability (i.e., both EOF-1 in Figure 2.4, and the first principal component would be enough to distinguish the two regimes) and they are the most geographically widespread clusters

- Atlantic Ridge (AR) confirms to be a "class" which is difficult to be completely decoupled from the other regimes. Again, the geographical configuration, and the visual results provided in Figure 2.8 support this idea

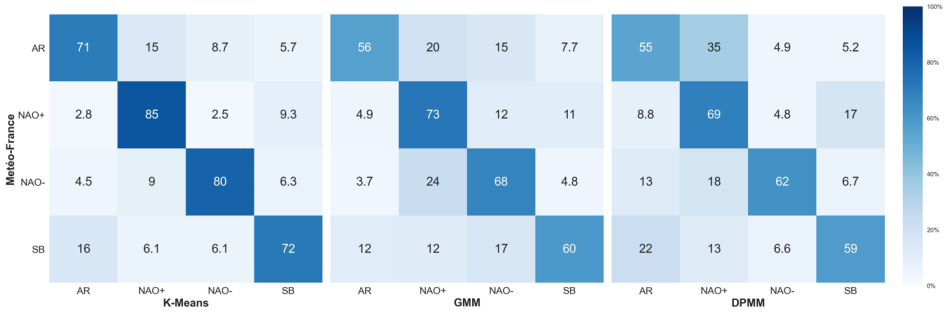


Figure 3.5: Confusion matrices between our 4 cluster models and Metéo-France historical predictions. The Mixture Models are inline with the agency's predictions. K-Means algorithm produces better scores, since it is aligned with the modeling choice of Meteo-France. From this illustration, it is evidenced how each model mostly strives with the Atlantic Ridge (AR) and Scandinavian Blocking (SB) configurations

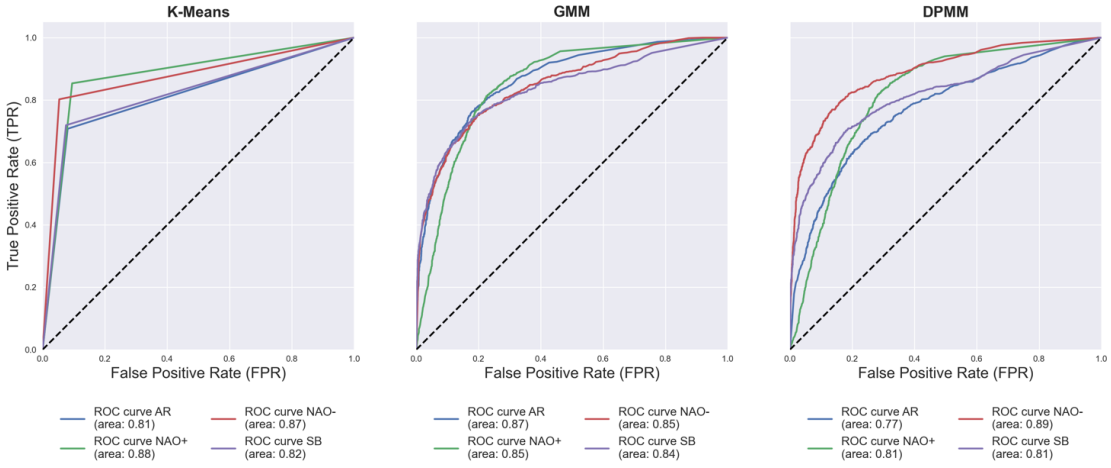
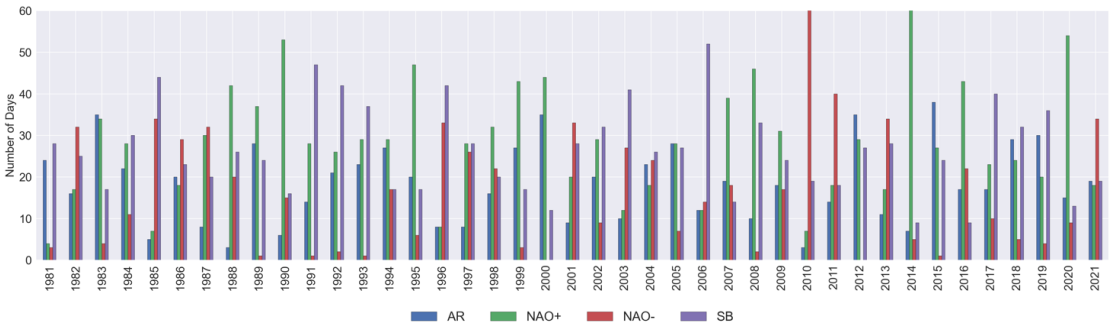
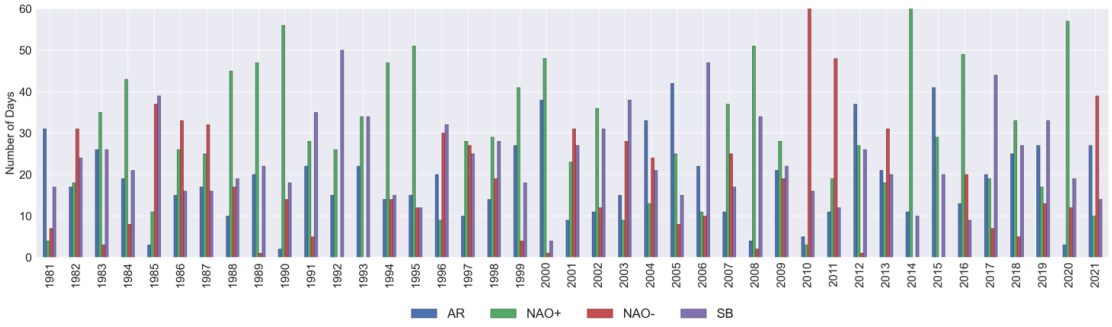


Figure 3.6: ROC curves of the three models: K-Means with PCA reduction (left), DPMM (center) and GMM (right) with VAE reduction

Further, we build the Receiver Operating Characteristic (ROC) curves of the three models. Figure 3.6 highlights the reduced performances of K-Means, compared to the probabilistic counterpart.
Since the Gaussian mixtures reach slighlty higher performances, the results obtained in the subsequent sections will be based on this modeling assumption.
Finally, zooming out to an historical overview of the incidence of regimes by winter (Figure 3.7), Météo-France and GMM share some similarities across the years, suggesting that our framework is coherent.
Again, we want to emphasize how our procedure tends to be superior under a computational efficiency point of view. We overcome the constraint of using ensemble models to track the chaotic nature of the regimes, simply focusing on a more complex and calibrated mathematical assumption.



(a) Metéo-France



(b) GMM

Figure 3.7: Historical counts of winter-days by regime

3.3.3 From Statistical To Physical Properties

The advantage of adopting probabilistic models is that clusters represent tractable distributions (Figure 3.8), which can be analyzed by means of analytically available statistics. We compute the Kullback-Leibler (KL) divergence between centroid

distributions of the different clusters, from which we are able to derive some interesting physical considerations.

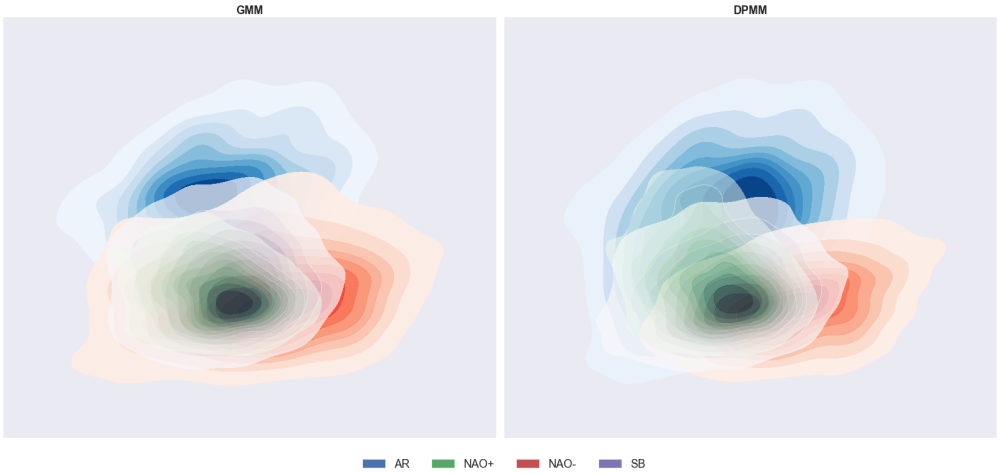


Figure 3.8: Visualization of the distributions of the clusters of the two mixture models along the two main modes of variability

Roughly speaking, the KL-divergence $\mathbb{KL}(p\|q)$ of a distribution p from a distribution q can be interpreted as the distance of an empirical distribution, obtained from independent samples drawn from q, to the distribution p. In other words, it maps how much samples from the target distribution q can be identified under the different probabilistic space of distribution p. It is then easy to understand that the opposite undergoes to a different relationship, and thus why the KL-divergence is not symmetric (i.e. $\mathbb{KL}(p\|q) \neq \mathbb{KL}(q\|p)$).

$\mathbb{KL}(p\|q)$	**AR**	**NAO+**	**NAO-**	**SB**
AR	0	**16**	5.7	5.9
NAO+	3.7	0	3.9	2.5
NAO-	5.7	**16**	0	**16**
SB	5.3	10	2.5	0

$\mathbb{KL}(p\|q)$	**AR**	**NAO+**	**NAO-**	**SB**
AR	0	**15**	5	8.5
NAO+	3.7	0	4.1	2.5
NAO-	6.2	**15**	0	**21**
SB	4.9	9.8	2.5	0

Table 3.6: KL-divergences of the regimes' distributions in GMM (left) and DPMM (right)

In the context of weather regimes, we firstly notice how the two Mixture models are quite in agreement, with minor differences only related to the magnitude of the divergences. Further, we identify that AR is difficult to be completely isolated, since its configurations result quite identifiable under all the other clusters. Indeed, the column of the AR regime presents considerable low scores under the other regimes'

distributions spaces. On the contrary, NAO+ patterns are well-separated from the other distributions since high scores are found for AR and NAO- configurations especially. This result is in line with physical considerations since SB, NAO- and AR are all known as "blocked" configurations (Scandinavian Blocking, Greenland Blocking and Atlantic Blocking), while NAO+ is a cyclonically dominated regime favouring the usual westerly flow over the European/Atlantic domain. We also want to highlight that $\mathbb{KL}(\text{NAO+}\|\text{SB}) = 2.5$ in both GMM and DPMM models, meaning that it is difficult to separate SB from NAO+. This situation reflects realistic physical scenarios. Indeed, several days with high probabilities for both SB and NAO+ are found.

3.3.4 Transitions

Ultimately, we are able to reconstruct an overview of the overall dynamics. We collect the historical transitions of the last 40 winters, as predicted by our models. To make the scenario more realistic, we consider a transition as valid only when a regime persists for at least 3 days, eventually disregarding single-day shifts interrupting the persistance of a regime. Again, the two models provide transition matrices in agreement, highlighting the aforementioned biases toward AR and NAO+ configurations. As expected, the tendency of these configurations is to persist in their state, thus justifying the higher values along the diagonal, and higher probabilities for larger-scale patterns like North Atlantic Oscillations, in accordance to (Büeler et al. 2021).
Empirically, the most usual transitions refer to the passages from NAO+ to SB, and from SB to NAO- (in agreement with (Vautard 1990)), as well as between AR and NAO+ in both the two directions. At the same time, the least frequent passages include the evolution from NAO+ to NAO- (Ferranti et al. 2018), and from NAO- to AR: all these aspects are clearly evident in our models.

From \ To	AR	NAO+	NAO-	SB
AR	77%	**9**%	8%	7%
NAO+	4%	88%	3%	**5**%
NAO-	3%	**6**%	86%	5%
SB	5%	**7**%	**7**%	80%

From \ To	AR	NAO+	NAO-	SB
AR	83%	**6**%	5%	**6**%
NAO+	4%	89%	2%	**6**%
NAO-	4%	**7**%	84%	5%
SB	7%	**8**%	4%	81%

Table 3.7: Transition Probabilities as predicted by GMM (left) and DPMM (right)

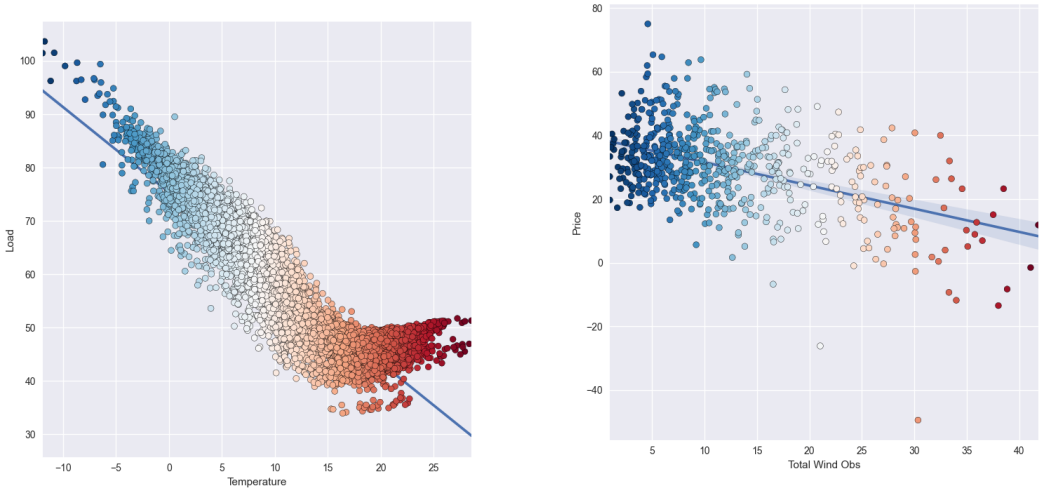
Chapter 4

Quantitative Analysis in the Energy domain

4.1 Energy-related variables

In the second part of our study, we employ our historical database of daily weather regimes to support a more refined quantification of correlated energy variables (Grams et al. 2017; van der Wiel et al. 2019a; van der Wiel et al. 2019b; Bloomfield et al. 2016).
Our focus is primarily on the EU-7 countries (France, Germany, United Kingdom, Italy, Spain, Belgium, Netherlands), mostly concerning Power demand, Wind/Solar production, and Hydro generation in the North pool (Scandinavia). Together, they account for almost the entirety of the power-trading operations in Europe, and our intent is to model more robustly the energy-related variables at the base of their strategies. To do so, as mentioned in Section 2.1.1, we need to plug in additional weather variables, which we know they tend to present very informative correlations with the energy counterparts. As an example, Figure 4.1 presents two very definite trends relating Temperature and Power Demand, and Wind Generation and Price, respectively. There is some evidence that weather regimes could potentially explain much of the underlying variability affecting energy markets. Especially for NAOs (De Felice et al. 2018; Jerez et al. 2013), which tend to be the predominant regimes during winter, a lot of the volatility of energy variables could be better addressed and quantified. Hence, in our second part we seek to answer the following questions:

- *Inter-regimes quantification*: how different regimes lead to different energetic configurations?
- *Intra-regimes quantification*: which effects mostly characterize a particular regime?



(a) Example of hockey-stick correlation between an additional weather variable (Temperature) and an energy-related variable (Power Demand). The correlation is strongly negative: at low temperatures heating systems drives most of the Power Load; at high temperatures, cooling systems are advocated for the increase in the correlation

(b) Example of relationship between the Total Wind Generation (GW) and Price (€/Mwh). It is interesting to notice an overall negative correlation, due to the production satisfying the market demand. Further, when the production overwhelms it (overproduction) it could happen that the price goes negative (energy providers are asked to buy the produced energy, and stock it)

Figure 4.1: Examples of relationships across weather and energy-related variables

- *Inter-counties quantification*: how the effects of a persistence of a regime vary with geographical position?

- *Intra-countries quantification*: how each country is subject to internal variability?

- *Extreme events*: which regimes most likely drive extreme events? How do they differ from one country to the other?

4.2 Quantifications

Energy variables like wind and solar production are known to be largely affected by meteorological configurations, with more unstable conditions favouring the former, while solar typically benefitting from dry environments. Besides, power demand correlates with temperature-related factors, again driven by these large scale systems. Hence, we first focus on the average configurations led by regimes in the different European countries.

Wind L.F. Anomaly	AR	NAO+	NAO-	SB
BE	0%	11%	-4%	-16%
ES	1%	2%	2%	-5%
FR	3%	6%	-1%	-10%
GE	5%	9%	-4%	-12%
IT	2%	-1%	0%	0%
NE	1%	9%	-5%	-14%
UK	2%	7%	-5%	-7%

Solar L.F. Anomaly	AR	NAO+	NAO-	SB
BE	-1%	-1%	0%	2%
ES	2%	0%	-2%	1%
FR	-1%	-1%	0%	3%
GE	-1%	-1%	0%	2%
IT	0%	0%	-1%	1%
NE	0%	-1%	0%	1%
UK	1%	0%	0%	0%

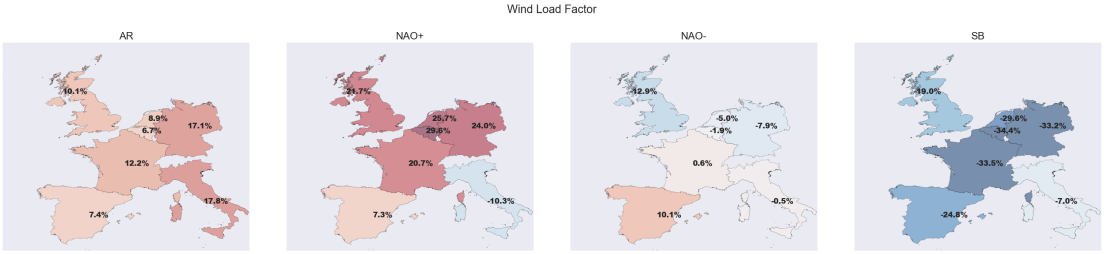
Load Bdays	AR	NAO+	NAO-	SB
BE	11,13	10,85	11,08	11,03
ES	31,62	30,78	30,90	31,31
FR	73,52	68,98	72,33	72,66
GE	67,30	65,63	66,85	66,71
IT	38,79	37,91	38,10	38,47
NE	14,32	14,13	14,36	14,28
UK	45,95	45,12	46,44	45,66

Figure 4.2: Wind Load Factor Deviation (%) (left), Solar Load Factor Deviation (%) (center), Power Demand (GW) during business days (right) under the different regimes for EU-7 countries

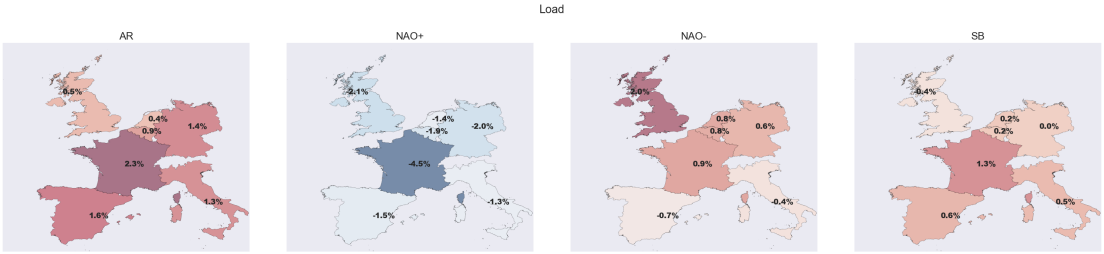
The results we obtain are reported by Figures 4.2, 4.3, and show very distinct behaviors that can be summarized as:

- Southern European countries (Italy and Spain) generally undergo to different setups, having milder conditions than Northern and Central countries

- Wind: the generally unstable conditions of NAO+ lead to much above normal productions of wind power plants, while Scandinavian Blocking (an usual dry regime) negatively affects how wind power plants can satisfy the power demand of a country. Further, Germany (GE) benefits from a regime like AR, generally leading to around normal winds. NAO-, instead, is a regime leading to cold temperatures and relatively dry meteorological setups. It is thus justified to have evidence of below normal productions



(a) Wind Load Factor Deviation (%) under the different regimes for EU-7 countries: a geographic visualization



(b) Power Demand Deviation (%) under the different regimes for EU-7 countries: a geographic visualization

Figure 4.3: Geographic visualization of the regimes and countries relationships of the Wind Load Factor and Power Demand variables

- Solar: being almost mutually exclusive with wind, data reflects an opposite situation. Here, SB is in average the regime under which production is boosted, whereas instead NAO+ ranks as the worst configuration. However, it is worth to notice that the magnitude of the deviations is much lower, clearly denoting the minor variability of this energy source

- Load: we analyze power demand on business days, since the effects of the industrial sectors can be taken into account. The statistics explicit an empirical trend widely discussed in meteorology: the presence of high wind and unstable conditions is usually associated with warmer temperatures. Indeed, NAO+ clearly reduces the overall power demand of a country, whose only sources of variability during weekdays are the private energy consumptions (in winter, heating systems). AR, NAO- and SB are instead similarly cold during winter, with the exception of southern countries under NAO- benefitting from more temperate climates

Collecting this fine-grained information is of uttermost importance, since it implicitly hides financial strategies in the energy markets. Recalling Figure 1.2, electricity prices are largely affected by how much demand can be satisfied by low-cost energy sources. Hence, we could expect that, under NAO+, extremely positive wind productions would likely lower the overall price, eventually allowing to keep gas and coal plants off. Instead, SB, although being favourable for Solar, should lead to higher price values due to the lack of wind.

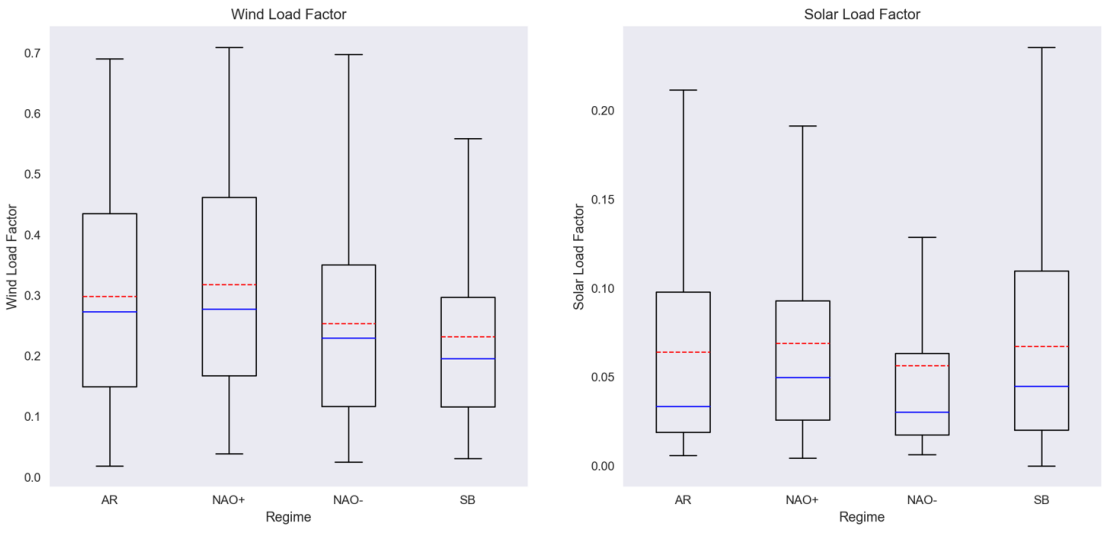


Figure 4.4: Example of historical distributions of the Wind and Solar Load Factor energy variables, isolated to an individual country (Germany)

As a next step, we recreate the distributions of many energy variables in the context of individual countries, to understand their statistical properties, and to assess a more faithful quantification. We are able to do this, considering the aforementioned historical series of measurements on the grid by the respective national operators. As an example, Figure 4.4 shows some insight into the distribution of the utilization of windy and solar power plants in Germany. Again, we are able to establish a bound on the maximum expected production from the two energy sources, by just considering what kind of events happened in history, and what was their intensity. For example, NAO+ and AR in Germany are comparable in their distribution, while NAO-, while largely variable, is generally less productive; SB is instead limited even in those days presenting an outlier behavior, suggesting a bearish trend in wind production. For the Solar counterpart, as outlined previously, variability is much more contained for most of the regimes, while clearly evidencing the benefits brought by the Blocking regime.

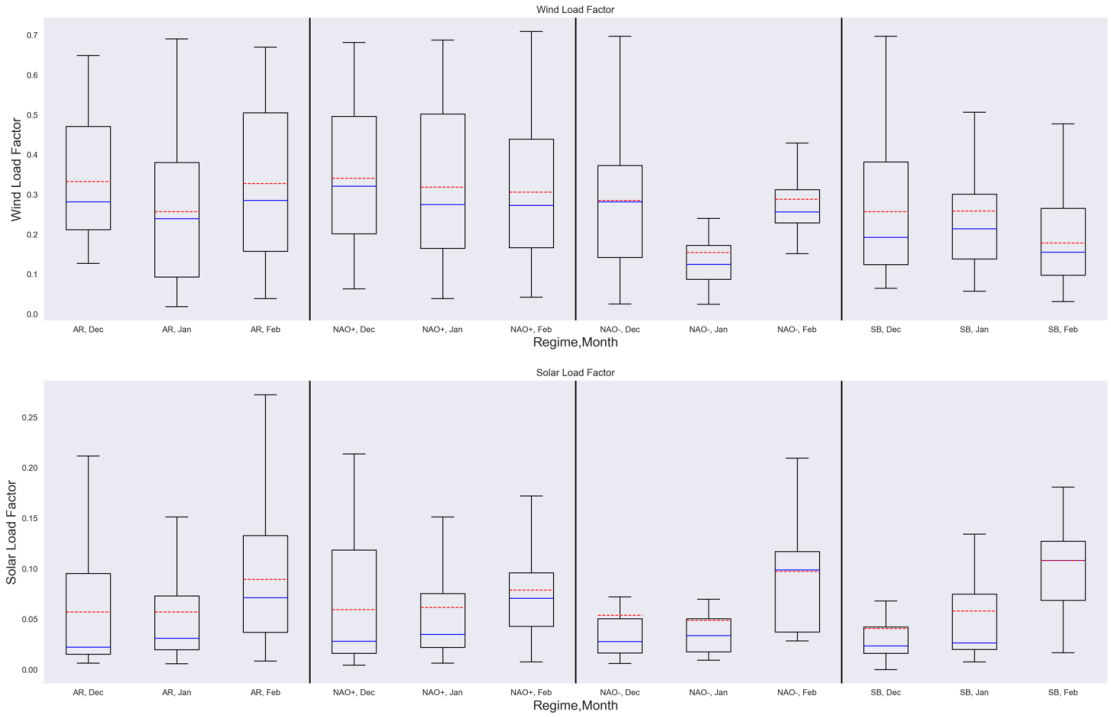


Figure 4.5: Example of monthly historical distributions of the Wind and Solar Load Factor energy variables, isolated to an individual country (Germany)

Diving deeper at a monthly level (Figure 4.5), we notice how January has a negative impact on the wind production under the NAOs and AR regimes. The explanation to this event can be found in an atmospheric phenomenon, usually happening at the beginning of the month: the Polar Vortex disruption. Its frequency

is not yearly, but when it occurs, wind events are typically milder than usual, and NAO- becomes the prominent regime for longer-than-average periods.
The only noticeable effect of Solar is instead the progressive increasing of the utilization factor as moving out from the winter season.
Finally, we perform a thorough analysis of extreme meteorological events in relation to weather regimes. We define an event as "extreme" in all those cases it falls beyond the 98th percentile of the distribution. Knowing which regimes are more likely to create disruptions is of paramount importance in that it allows to predict huge volatilities in energy markets. The results we collect are reported in Figure 4.6.

Wind L.F.	AR	NAO+	NAO-	SB
BE	9.4	63.0	23.2	4.5
ES	15.6	50.9	26.7	6.7
FR	25.3	48.7	22.8	3.2
GE	35.6	47.2	12.6	4.6
IT	30.8	26.9	21.5	20.8
NE	15.7	48.5	25.5	10.3
UK	33.2	45.3	14.3	7.2

Load	AR	NAO+	NAO-	SB
BE	10.9	11.6	51.8	25.8
ES	23.7	12.4	32.9	31.0
FR	12.7	6.7	47.1	33.6
GE	8.4	9.2	44.7	37.7
IT	17.3	11.7	33.7	37.4
NE	7.4	14.2	48.9	29.5
UK	7.2	16.3	55.3	21.2

Figure 4.6: Occurrence (%) of extreme events in EU-7 countries in the four different regimes for Wind (left) and Power Demand (right)

Chapter 5

Sub-seasonal Forecasts

The final step of the pipeline consists in transposing the quantification step into a medium-term outlook. This is achieved by injecting the knowledge derived from the product presented in Section 2.1.3, thereby calibrating it by means of our predicted distributions. As previously mentioned, one limitation of such forecasts are in the clustering assumption they rely on. Indeed, ECMWF (Ferranti et al. 2011) adopts a 5-clusters ensemble model, not encompassing the entirety of the variability of weather regimes, and thus needing the presence of an unknown cluster to take into account noisier assignments. Since our application strictly requires a 4-clusters scenario, we approximate the fifth missing distribution by assigning probability mass to the other distributions, proportionally to the assigned weight each regime has. Of course, whether the "Unknown" cluster is predicted with a 100% probability, we keep it uninformative assigning uniform probabilities to all the clusters' distributions.
The calibration step is thus obtained by computing a weighted sum of the four distributions. Although for some energy variables the distributions referred to the regimes are really close, we observe that minimum shifts in the mean and standard deviation often translate into typically different outlooks for the predicted window of 45 days.
In our analysis, we are also able to take into account known biases of these forecasts. Figure 5.1 reports the tendency of these statistical models to favour Zonal flow in the long-term, while striving with the NAO- signals. Indeed, as a general pattern, the average probabilities assigned to the regimes should gradually converge to an uninformative fashion, as the forecasted step is farther in time. However, the NAO+ regime's presence (conceptually equivalent to Zonal flow) increases in importance. Unfortunately, given the novelty of the product, the chart is only referred to the last winter predictions, thus being biased on the events which charactrized it. We clearly see in the initial steps the overall high probabilities assigned to NAO-, reflecting correctly the tendency of the regimes during that season. The peak in

Scandinavian Blocking is instead difficult to be justified, given that its occurrence in Winter 2020 resulted to be below normal.

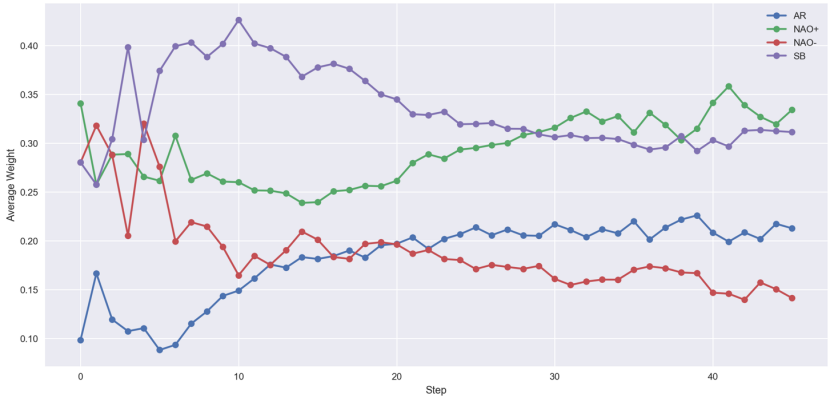


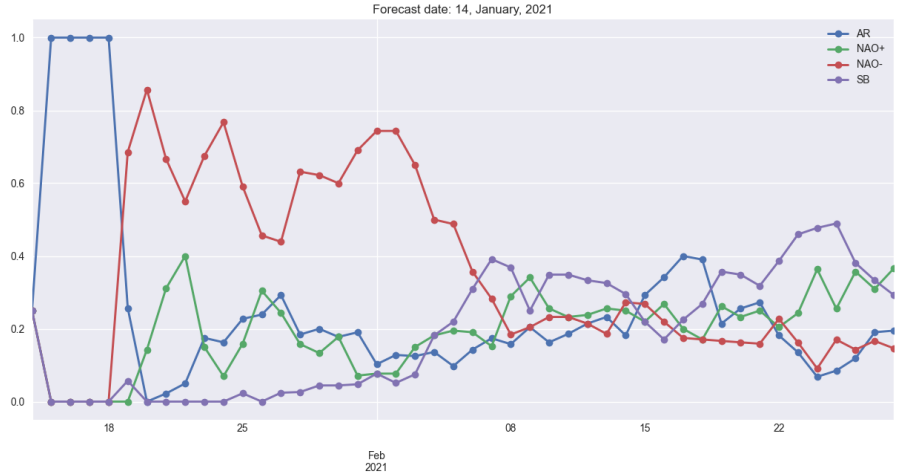
Figure 5.1: Average weights of the sub-seasonal forecasts of ECMWF by regime

We adopt this analysis as a prior belief, mainly to establish some high-level criteria to identify whether some signals could be considered enough strong, and thus potentially accurate with some weeks of anticipation.

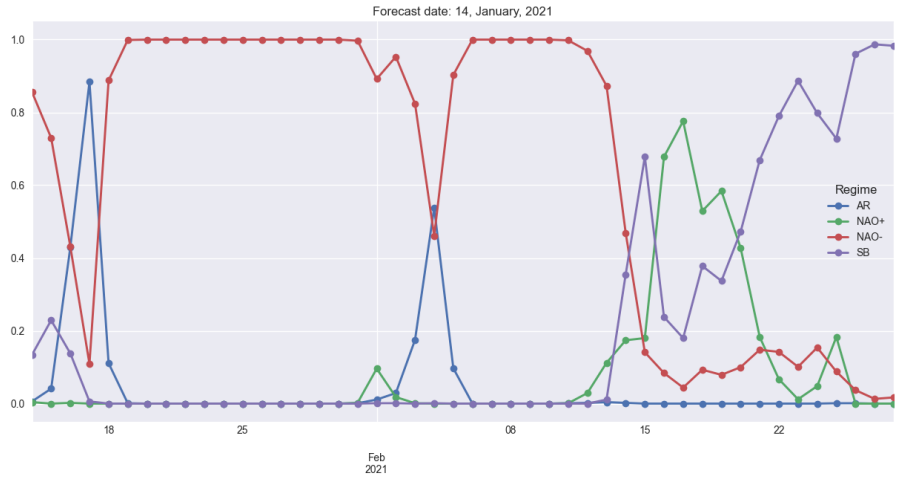
5.1 Case Study: when the forecast is accurate

As a first case study, we consider the forecast released on 14th January 2021 (Figure 5.2a). The sub-seasonal model is able to detect the main trends, if compared to the a-posteriori probabilities deriving from our GMM Model (Figure 5.2b). The initial period is characterized by a very strong AR signal, replaced one week later by a long period of NAO- persistence. The forecast is also able to identify an emerging signal of Scandinavian Blocking, which effectively occurred at the end of the period.

The end-product we are able to present is illustrated in Figure 5.3. The chart is referred to the anomaly in the Wind production of power in Germany along the forecasted period. As the Quantification step highlights, the initial presence of AR is particularly favourable for above-normal Wind generation. However, the scope of this tool is to effectively address the trend beyond the first few days, where short-term forecasts would surely be more accurate, especially when referred to the expected magnitude. The presence of NAO- and the subsequent transition to SB are two main drivers for below normal anomalies, which are clearly evidenced by the chart, and correspondingly benchmarked by actual data: in the window between the end of January and the first half of February the production in Germany revelead to be 4GW below normal; during the Blocking persistence was instead 3GW under the average scenario.



(a) Sub-seasonal Forecast weights of 14th January 2021



(b) Clustering Probabilities of the GMM Model

Rather than exactly quantifying the shift from the average case, we intend to identify the trend a particular energy variable has in this window of opportunity. For this reason, even when the forecast proves to be performant, it is rarely close to the true observed value.

5.2 Case Study: when the forecast is inaccurate

The strict dependency on the underlying forecasts exposes the main point of weakness of this final step. A totally incorrect forecast, albeit typically rare, is a

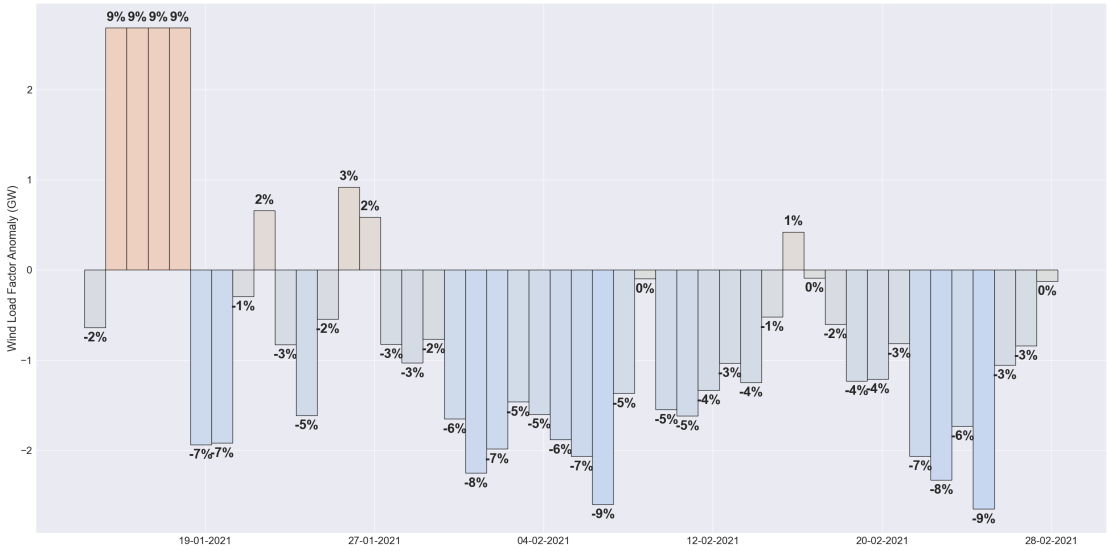
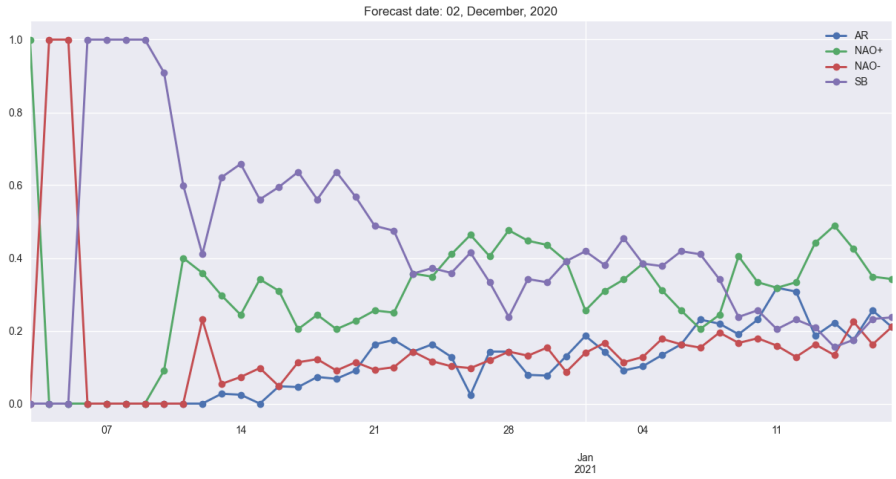
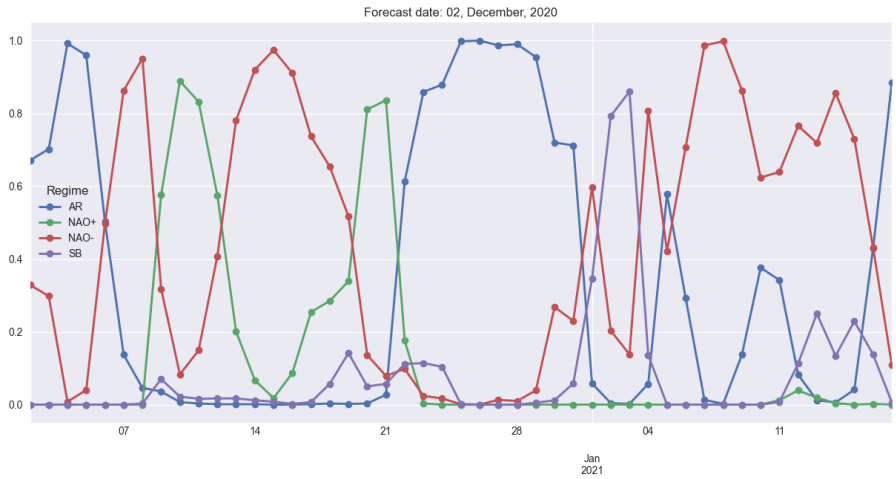


Figure 5.3: Quantification of the Sub-seasonal forecast on the Wind Load Factor Anomaly in Germany

driver of an as much innacurate quantification. We thus propose a scenario where we identify this occurrence, corresponding to 2nd December 2020. Comparing the results between our model's prediction and the product of ECMWF (Figures 5.4a and 5.4b), the sub-seasonal model is not aligned with the regimes effectively occurred over that period, whereas they are more closely matched by the clustering results evaluated posteriorly.



(a) Sub-seasonal Forecast weights of 2nd December 2020



(b) Clustering Probabilities of the GMM Model

The calibration and quantification steps in the 45 days window systematically reflect this error. The snapshot is referred to the wind generation occurring in Spain. The below normal generation interesting the vast majority of the forecast

turned out to be in reality much above normal, as expected by the NAO- persistence proving to set adequate conditions for extreme wind events in Spain.

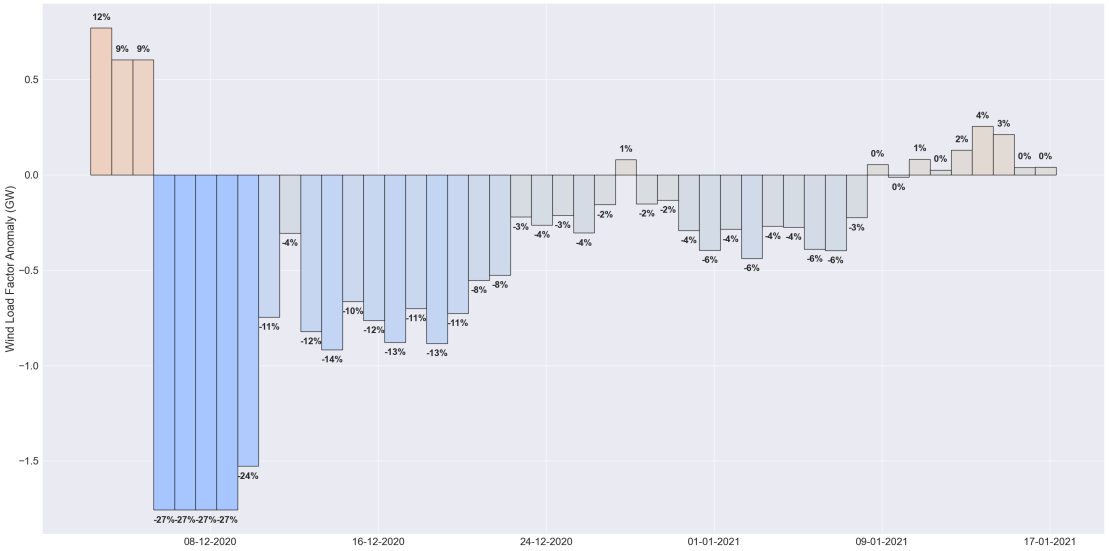


Figure 5.5: Quantification of the Sub-seasonal forecast on the Wind Load Factor Anomaly in Spain

Chapter 6

Conclusions

The study has thoroughly analyzed the dynamics of weather regimes during the winter season in the North Atlantic-European zone. Starting from raw meteorological data, we built an historical dataset of daily weather regimes, based on a more sophisticated modeling assumption. Consequently, it has allowed us to perform a finer quantification both in the meteorological and energy domain, evidencing some neat trends under the different regimes' distributions. Such knowledge entails an instrument to better address and interpret weather impact in the context of energy markets, clearly representing one of the "fundamentals" used in the trading strategies.

As a last step, we have leveraged a new product in the weather community, possibly opening new future scenarios in the forecasting domain. The span of medium-term forecasts, combined with the predictability of the regimes configurations, clearly play a pivotal role in determining a window of opportunity, where some trends can be spotted up to 4-5 weeks with anticipation. Of course, the current limitation of the product, both in terms of accuracy and recent availability, are two key limiting factors for a continuous deployment and assessment of the performances, therefore of any possibility to remove some innermost biases of the underlying forecasting model.

In terms of contributions of our study, we have proved that different modeling assumptions can be pursued. If the aim is to simply resort to the centroid configurations, the standard techniques (PCA and K-Means) presented in literature are enough for it. However, if the task is extended to a fine-grained quantification, it is essential to switch the modeling paradigm to a probabilistic one. The overall global results remain inline with some references in literature, however it is interesting to derive some secondary knowledge such as the evolution of the patterns, and the statistical properties of the distributions matching some expected physical assumptions. Resorting to the approximation power of neural networks have given us more flexibility, drastically reducing the dimensionality of the feature space,

while improving the separation of data points. The overall interpretability is enhanced too, given the optimization process leading to disentangled representations of the attributes. Thereby, the adoption of Mixture Models has led to a major robustness of the framework, as well as to an improved consistency of the clustering assignments.

Along this line, future works should then focus on some experimental and critical points we have encountered during the process. Firstly, the adoption of VAEs for the dimensionality reduction step could be helpful to consider several snapshots of different atmospheric levels. Indeed, differently than PCA, Variational AutoEncoders can receive as input multiple-channels images, over which to perform the encoding step. In this context, instead of feeding a single anomaly map of the Geopotential at 500hPa (van der Wiel et al. 2019a; Cassou et al. 2011; Grams et al. 2017; Vrac et al. 2007), we could think to aggregate multiple pressure levels such as 700/850 hPa (Vautard 1990; Hertig et al. 2014; Vrac et al. 2007), or sea level pressure (Meteo-France n.d.; Vrac et al. 2007), thus better characterizing the datapoints and overcoming any strict choice in terms of which weather variable to consider.

Further, as the results on the probabilistic models show, a slight difference was spotted on the expected frequencies for AR and NAO- regimes. Whether this is a bias of the models, or an evidence of more faithful results has to be investigated. Then, other probabilistic modeling choices could be pursued (Stan et al. 2017), possibly leading to major performances.

On the quantification side, on the hand the limited data availability on the true measurements, on the other hand the sometimes misaligned nature of the synthetic data, represent two key limiting factors. In our experiments, we resort to techniques like quantile-quantile (QQ) mapping to "re-align" the synthetic data distribution to the true observation, however with poor results. Some efforts on this side could then be productive for a still more faithful analysis.

Finally, the forecasting tool we develop, as we mention in Chapter 5, is largely dependent on the precision of the underlying product from ECMWF. Surely, this represents the step with possibilities of larger improvement. Having access to raw forecasts of geopotential maps would represent the best possible scenario, ensuring the consistency with all the methods developed inside the pipeline. Further, it would allow to test several clusters conformation, differing by the number of regimes adopted, not mandatorily sticking with the theoretical ones. Alternatively, spotting some innermost biases the ECMWF model has, would translate into accentuating some longer-term signals, esepcially when they are uniformly mixed between each other. To accomplish this, however, several years of data should be necessary, thus not representing a feasible short-term solution.

Bibliography

Hertig, Elke and Jucundus Jacobeit (2014). «Variability of weather regimes in the North Atlantic-European area: past and future». In: *Atmospheric Science Letters* 15.4, pp. 314–320. DOI: https://doi.org/10.1002/asl2.505. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl2.505. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl2.505 (cit. on pp. 7, 60).

Vautard, R. (1990). «Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors». In: *Monthly Weather Review* 118.10, pp. 2056–2081. DOI: 10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/118/10/1520-0493_1990_118_2056_mwrotn_2_0_co_2.xml (cit. on pp. 7, 45, 60).

Meteo-France (n.d.). *Climate monitoring - Daily monitoring of weather regimes*. URL: http://seasonal.meteo.fr/ (cit. on pp. 7, 13, 36, 37, 40, 60).

Vrac, Mathieu, Katharine Hayhoe, and Michael Stein (Apr. 2007). «Identification of intermodal comparison of seasonal circulation patterns over North America». In: *International Journal of Climatology* 27, pp. 603–620. DOI: 10.1002/joc.1422 (cit. on pp. 7, 25, 60).

Robertson, Andrew W. and Michael Ghil (1999). «Large-Scale Weather Regimes and Local Climate over the Western United States». In: *Journal of Climate* 12.6, pp. 1796–1813. DOI: 10.1175/1520-0442(1999)012<1796:LSWRAL>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/clim/12/6/1520-0442_1999_012_1796_lswral_2.0.co_2.xml (cit. on p. 7).

Bruyère, Cindy, Chainarong Raktham, J Done, J Kreasuwun, J Thongbai, and Wonchai Promnopas (Jan. 2016). «Major weather regime changes over Southeast Asia in a near-term future scenario». In: *Climate Research* 72. DOI: 10.3354/cr01442 (cit. on p. 7).

Cassou, Christophe, Laurent Terray, and Adam Phillips (Aug. 2005). «Tropical Atlantic Influence on European Heat Waves». In: *Journal of Climate - J CLIMATE* 18, pp. 2805–2811. DOI: 10.1175/JCLI3506.1 (cit. on pp. 7, 9, 25).

Weather and Climate at Reading (n.d.). *Weather Regimens Visualization*. URL: http://blogs.reading.ac.uk/weather-and-climate-at-reading/2020/recent-progress-in-simulating-north-atlantic-weather-regimes/ (cit. on p. 8).

Straus, David, Susanna Corti, and Franco Molteni (May 2007). «Circulation Regimes: Chaotic Variability versus SST-Forced Predictability». In: *Journal of Climate* 20, pp. 2251–. DOI: 10.1175/JCLI4070.1 (cit. on p. 9).

Grams, Christian, Remo Beerli, Stefan Pfenninger, Iain Staffell, and Heini Wernli (July 2017). «Balancing Europe's wind-power output through spatial deployment informed by weather regimes». In: *Nature Climate Change* 7. DOI: 10.1038/nclimate3338 (cit. on pp. 9, 12, 25, 36, 40, 46, 60).

Falkena, Swinda K. J., J. D. Wiljes, A. Weisheimer, and T. Shepherd (2020). «Revisiting the Identification of Winterftime Atmospheric Circulation Regimes in the Euro-Atlantic Sector». In: (cit. on pp. 9, 12, 25, 35).

Cassou, Christophe, Marie Minvielle, Laurent Terray, and Claire Périgaud (Jan. 2011). «A statistical-dynamical scheme for reconstructing ocean forcing in the Atlantic. Part I: weather regimes as predictors for ocean surface variables». In: *Climate Dynamics* 36.1-2, pp. 19–39. DOI: 10.1007/s00382-010-0781-7 (cit. on pp. 9, 12, 25, 36, 37, 40, 60).

van der Wiel, K., Hannah C Bloomfield, Robert W Lee, Laurens P Stoop, Russell Blackport, James A Screen, and Frank M Selten (Sept. 2019a). «The influence of weather regimes on European renewable energy production and demand». In: *Environmental Research Letters* 14.9, p. 094010. DOI: 10.1088/1748-9326/ab38d3. URL: https://doi.org/10.1088/1748-9326/ab38d3 (cit. on pp. 9, 12, 25, 36, 37, 40, 46, 60).

Climate Prediction Center (n.d.). *Climate Prediction Center - Monthly North Atlantic Oscillation Index* (cit. on pp. 9, 17).

Store, Copernicus - Climate Data (n.d.[a]). *Climate Data Store*. URL: https://cds.climate.copernicus.eu/cdsapp%5C#!/home (cit. on p. 11).

Hersbach, Hans et al. (2020). «The ERA5 global reanalysis». In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049 (cit. on p. 12).

Store, Copernicus - Climate Data (n.d.[b]). *ERA5 hourly data on pressure levels from 1979 to present*. URL: https://cds.climate.copernicus.eu/cdsapp%5C#!/dataset/reanalysis-era5-pressure-levels (cit. on p. 12).

Ferranti, Laura and S. Corti (2011). «New clustering products». In: (127), pp. 6–11. DOI: 10.21957/lr3bcise. URL: https://www.ecmwf.int/node/17442 (cit. on pp. 14, 53).

Hannachi, Abdel (Apr. 2004). «A primer for EOF analysis of climate data». In: (cit. on p. 15).

Elbattah, Mahmoud, Colm Loughnane, Jean-Luc Guérin, Romuald Carette, Federica Cilia, and Gilles Dequen (2021). «Variational Autoencoder for Image-Based

Augmentation of Eye-Tracking Data». In: *Journal of Imaging* 7.5. ISSN: 2313-433X. DOI: 10.3390/jimaging7050083. URL: https://www.mdpi.com/2313-433X/7/5/83 (cit. on p. 19).

Saenz, J., N. Lubbers, and Nathan Urban (Aug. 2018). «Dimensionality-Reduction of Climate Data using Deep Autoencoders». In: (cit. on p. 18).

Prasad, Vignesh, Dipanjan Das, and Brojeshwar Bhowmick (2020). «Variational Clustering: Leveraging Variational Autoencoders for Image Clustering». In: *CoRR* abs/2005.04613. arXiv: 2005.04613. URL: https://arxiv.org/abs/2005.04613 (cit. on p. 18).

LearnOpenCV (n.d.). *Variational Autoencoder in TensorFlow*. URL: https://learnopencv.com/variational-autoencoder-in-tensorflow/ (cit. on p. 19).

Locatello, Francesco, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem (2018). «Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations». In: *CoRR* abs/1811.12359. arXiv: 1811.12359. URL: http://arxiv.org/abs/1811.12359 (cit. on p. 21).

Locatello, Francesco, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem (2019a). «On the Fairness of Disentangled Representations». In: *CoRR* abs/1905.13662. arXiv: 1905.13662. URL: http://arxiv.org/abs/1905.13662 (cit. on p. 21).

Locatello, Francesco, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem (2019b). «Disentangling Factors of Variation Using Few Labels». In: *CoRR* abs/1905.01258. arXiv: 1905.01258. URL: http://arxiv.org/abs/1905.01258 (cit. on p. 21).

Locatello, Francesco, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen (July 2020). «Weakly-Supervised Disentanglement Without Compromises». In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6348–6359. URL: http://proceedings.mlr.press/v119/locatello20a.html (cit. on p. 21).

Träuble, Frederik, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer (July 2021). «On Disentangled Representations Learned from Correlated Data». In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 10401–10412. URL: http://proceedings.mlr.press/v139/trauble21a.html (cit. on p. 21).

Dittadi, Andrea, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf (2020). «On the Transfer of Disentangled Representations in Realistic Settings». In: *CoRR*

abs/2010.14407. arXiv: 2010.14407. URL: https://arxiv.org/abs/2010.14407 (cit. on p. 21).

Mita, Graziano, Maurizio Filippone, and Pietro Michiardi (2021). *An Identifiable Double VAE For Disentangled Representations*. arXiv: 2010.09360 [cs.LG] (cit. on p. 21).

Higgins, I., L. Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner (2017). «beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework». In: *ICLR* (cit. on p. 22).

Rybkin, Oleh, Kostas Daniilidis, and Sergey Levine (2020). *Simple and Effective VAE Training with Calibrated Decoders* (cit. on p. 22).

van der Wiel, K., L.P. Stoop, B.R.H. van Zuijlen, R. Blackport, M.A. van den Broek, and F.M. Selten (2019b). «Meteorological conditions leading to extreme low variable renewable energy production and extreme high energy shortfall». In: *Renewable and Sustainable Energy Reviews* 111, pp. 261–275. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2019.04.065. URL: https://www.sciencedirect.com/science/article/pii/S1364032119302862 (cit. on pp. 25, 46).

Bloomfield, D J Brayshaw, L C Shaffrey, P J Coker, and H E Thornton (Dec. 2016). «Quantifying the increasing sensitivity of power systems to climate variability». In: *Environmental Research Letters* 11.12, p. 124025. DOI: 10.1088/1748-9326/11/12/124025. URL: https://doi.org/10.1088/1748-9326/11/12/124025 (cit. on pp. 25, 46).

Smyth, Padhraic, Kayo Ide, and Michael Ghil (1999). «Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering». In: *Journal of the Atmospheric Sciences* 56.21, pp. 3704–3723. DOI: 10.1175/1520-0469(1999)056<3704:MRINHH>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/56/21/1520-0469_1999_056_3704_mrinhh_2.0.co_2.xml (cit. on p. 25).

Stan, Cristiana, David M. Straus, Jorgen S. Frederiksen, Hai Lin, Eric D. Maloney, and Courtney Schumacher (2017). «Review of Tropical-Extratropical Teleconnections on Intraseasonal Time Scales». In: *Reviews of Geophysics* 55.4, pp. 902–937. DOI: https://doi.org/10.1002/2016RG000538. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016RG000538. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016RG000538 (cit. on pp. 25, 60).

Blei, David M. and Michael I. Jordan (2006). «Variational inference for Dirichlet process mixtures». In: *Bayesian Analysis* 1.1, pp. 121–143. DOI: 10.1214/06-BA104. URL: https://doi.org/10.1214/06-BA104 (cit. on p. 30).

Luo, Dehai, Jing Cha, and Steven B. Feldstein (2012). «Weather Regime Transitions and the Interannual Variability of the North Atlantic Oscillation. Part I: A Likely Connection». In: *Journal of the Atmospheric Sciences* 69.8, pp. 2329–2346. DOI:

10.1175/JAS-D-11-0289.1. URL: https://journals.ametsoc.org/view/journals/atsc/69/8/jas-d-11-0289.1.xml (cit. on pp. 36, 37, 40).

Büeler, Dominik, L. Ferranti, Linus Magnusson, Julian Quinting, and Christian Grams (Sept. 2021). «Year-round sub-seasonal forecast skill for Atlantic-European weather regimes». In: *Quarterly Journal of the Royal Meteorological Society*. DOI: 10.1002/qj.4178 (cit. on p. 45).

Ferranti, Laura, Linus Magnusson, Frédéric Vitart, and David S. Richardson (2018). «How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe?» In: *Quarterly Journal of the Royal Meteorological Society* 144.715, pp. 1788–1802. DOI: https://doi.org/10.1002/qj.3341. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3341. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3341 (cit. on p. 45).

De Felice, Matteo, Laurent Dubus, Emma Suckling, and Alberto Troccoli (July 2018). «The impact of the North Atlantic Oscillation on European hydro-power generation». In: DOI: 10.31223/osf.io/8sntx (cit. on p. 46).

Jerez, Sonia, Ricardo Trigo, Sergio Vicente-Serrano, D. Pozo-Vazquez, Raquel Lorente-Plazas, Jorge Lorenzo-Lacruz, Francisco Santos-Alamillos, and J. Montávez (Oct. 2013). «The Impact of the North Atlantic Oscillation on Renewable Energy Resources in Southwestern Europe». In: *Journal of Applied Meteorology and Climatology* 52, pp. 2204–2225. DOI: 10.1175/JAMC-D-12-0257.1 (cit. on p. 46).