FlexPlan

Advanced methodology and tools taking advantage of storage and FLEXibility in transmission and distribution grid PLANning

Monte Carlo scenario generation and reduction

D1.1

Distribution Level	PU
Responsible Partner	TU Dortmund University (TUDO)
Checked by WP leader	Hakan Ergun (KU Leuven) – WP1 Leader Date: 09/12/2020
Verified by the appointed Reviewers	Maxime Hanot (N-SIDE) – Date: 16/11/2020 Iver Bakken Sperstad (SINTEF) – Date: 16/11/2020
Approved by Project Coordinator	Gianluigi Migliavacca (RSE) Date: 10/12/2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863819

Issue Record

Planned delivery date	01/07/2020
Actual date of delivery	10/12/2020
Status and version	v1.0

Version	Date	Author(s)	Notes
0.1	15.04.2020	Tobias Patzwald (TUDO), Reinhilde D'hulst (VITO)	First draft of the deliverable containing the intended structure of the task, the exhaust and definitions of the individual subtasks.
0.2	20.05.2020	Tobias Patzwald (TUDO), Reinhilde D'hulst (VITO)	Second draft of the Deliverable with a first version of a content elaboration for the methods presented.
0.3	31.07.2020	Björn Matthes (TUDO)	Revised structure and content fundamentally.
0.4	01.09.2020	Björn Matthes (TUDO)	Rewrite Introduction, State of the art, and Methodology from scratch.
0.5	30.09.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	New draft of the deliverable. Including state of the art, scenario generation and reduction approach.
0.6	05.10.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	First draft of complete document's content.
0.7	13.10.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	Integrated feedback of WP leader.
0.8	23.10.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	Final draft for internal revision.
0.8.1	16.11.2020	Hakan Ergun (KU Leuven), Maxime Hanot (N-SIDE), Iver Bakken Sperstad (SINTEF)	Internal revision completed.
0.9	01.12.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	Integrated feedback from internal reviewers.
1.0	04.12.2020	Björn Matthes (TUDO), Reinhilde D'hulst (VITO)	Integrated feedback of WP leader.

About FlexPlan

The FlexPlan project aims at establishing a new grid planning methodology considering the opportunity to introduce new storage and flexibility resources in electricity transmission and distribution grids as an alternative to building new grid elements. This is in line with the goals and principles of the new EC package *Clean Energy for all Europeans,* which emphasizes the potential usage of flexibility sources in the phases of grid planning and operation as alternative to grid expansion. In sight of this, FlexPlan creates a new innovative grid planning tool whose ambition is to go beyond the state of the art of planning methodologies, by including the following innovative features: integrated T&D planning, full inclusion of environmental analysis, probabilistic contingency methodologies replacing the N-1 criterion as well as optimal planning decision over several decades. However, FlexPlan is not limited to building a new tool but it also uses it to analyse six regional cases covering nearly the whole European continent, aimed at demonstrating the application of the tool on real scenarios as well as at casting a view on grid planning in Europe till 2050. In this way, the FlexPlan project tries to answer the question of which role flexibility could play and how its usage can contribute to reduce planning investments yet maintaining (at least) the current system security levels. The project ends up formulating guidelines for regulators and for the planning offices of TSOs and DSOs. The consortium includes three European TSOs, one of the most important European DSO group, several R&D companies and universities from 8 European Countries (among which the Italian RSE acting as project coordinator) and N-SIDE, the developer of the European market coupling platform EUPHEMIA.

Partners

Table of Contents

List of Abbreviations and Acronyms

Abbreviation/Acronym	Meaning
AC	Alternating Current
BRL	Boland-Ridley-Lauret
CAPEX	Capital Expenditure
DC	Direct Current
DWD	German Weather Service
ECMWF	European Centre for Medium-Range Weather Forecasts
E-HYPE	European Hydrological Predictions for the Environment
ENTSO-E	European Network of Transmission System Operators for Electricity
GMAO	Global Modeling and Assimilation Office
GSEE	Global Solar Energy Estimator
JMA	Japan Meteorological Agency
LIMES-EU	Long-term Investment Model for the Electricity Sector
MC	Monte Carlo
MILES	Model of International Energy Systems
NASA	National Aeronautics and Space Administration
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NUTS	Nomenclature of Territorial Units for Statistics
OPEX	Operational Expenditure
Pan-EU	Pan-European
PCA	Principal Component Analysis
PV	Photovoltaic
RC	Regional Case
RoR	Run of River
SMHI	Swedish Meteorological and Hydrological Institute
TIMES	The Integrated MARKAL-EFOM System
TSG	Time Series Generation
TSO	Transmission System Operator
TYNDP	Ten-Year Network Development Plan
vRES	Variable renewable energy sources
VWF	Virtual Wind Farm

Executive Summary

This document describes the Monte Carlo (MC) scenario generation and reduction methodology developed within the FlexPlan project. The developed methodology is used to first generate a large variety nodal generation and demand scenarios in terms of hourly time series, respectively. The generated time series are further reduced to a representative set of time series which are used as input for the advanced planning tool implemented within FlexPlan.

The intermittent generation from variable renewable energy sources, the generation of hydro power plants and the electricity demand are considered as stochastic inputs with respect to the grid expansion planning problem. FlexPlan's advanced planning approach incorporates storage and demand flexibility as alternatives to classical grid expansion. As such, time series data is required as input for the advanced planning tool in order to accurately represent the intertemporal constraints linked to the operational characteristics of demand flexibility and storage. Thus, the developed scenario generation and reduction methodology needs to provide hourly time series for all stochastic inputs of the planning problem at hand.

The developed methodology and implemented prototype consist of the following building blocks:

- A geographic reference system based on approximately 290 locations in Europe for time series data
- A database, containing meteorological and hydrological information for 40 years
- A time series generator for wind, solar and hydropower generation sampling
- A time series generator for demand sampling
- A method to model temporal and spatial correlations of the aforementioned time series
- A methodology to reduce a huge amount of operational scenarios formed by the aforementioned generation and load time series to a representative set

To develop the proposed methodology, significant work has been done on the modelling of temporal and spatial correlations between the stochastic inputs influencing expansion planning, on the identification of sufficient data sources (considering the limited availability of publicly available high quality historic power system data) and on finding a compromise between the required level of detail and the computational complexity by modelling uncertainties in large scale expansion planning problems.

State-of-the art literature was reviewed with regard to the advantages and disadvantages of different modeling methods and their suitability to be used in the context of FlexPlan. Furthermore, the designed methods for scenario generation and reduction were analyzed and validated for a proof-of-concept test case.

1. Introduction

The transmission and distribution grid planning tool developed within FlexPlan aims at finding optimal grid reinforcement and expansion measures including storage and demand flexibility over several decades (2030-2040-2050) on a pan-European level and on regional levels. The future energy system analyzed with the FlexPlan tool is characterized by a number of long-term visions, describing possible developments of the energy system as well as divergent European energy policies.

The Pan-European scenarios developed in FlexPlan include macro-assumptions on the generation mix and the projected demand on a country level. Using the *Model of International Energy Systems (MILES)* [1] [2] [3] of TU Dortmund, these macro-assumptions are broken down to smaller regions. Thus, national installed capacities per generation and load category are spatially disaggregated to significantly smaller sub-regions in each country (see FlexPlan's D4.1 [4]). Furthermore, large scale technological trends with regard to generation technologies and the development of conventional power plants have been considered as part of the Pan-European macro scenarios adopted by FlexPlan.

The regionalized Pan-EU scenarios describe the evolution of the energy system on a level that is not sufficient to carry out detailed grid expansion studies. In order to determine the optimal grid expansion on a nodal and zonal level, the pan EU-scenarios have to be disaggregated to individual nodes in the transmission and distribution grids such that they can be used by the advanced planning tool. Furthermore, the dependency of non-dispatchable units, e.g. wind, solar and partly hydro power generators, on local climatic conditions have to be considered in order to obtain realistic results within the regional studies. Therefore, a scenario generation approach is used, determining the hourly power generation of non-dispatchable units considering uncertainties. To ensure ***maximum consistency*** between the macro-assumptions taken within the pan-EU scenarios and the detailed nodal operational states considered in the regional cases, the detailed sub-regional output of the *MILES* platform is used as input for the developed *Scenario Generation and Reduction Approach* presented in this deliverable.

1.1. Situation of this deliverable within the FlexPlan project

The conceptual block diagram depicted in Figure 1-1 visualizes FlexPlan's key research questions and the intended approach to tackle them. Central questions of the new planning tool's conceptual design are addressed in WP 1. Furthermore, technologies for flexibility provision are analyzed in depth in WP 2 utilizing a pre-processor tool to evaluate flexibility elements' characteristics, e.g. CAPEX, OPEX and expected lifetimes. The same pre-processing tool is utilized to provide indications on the sizing and location of storage candidates, as well as candidate connections that are finally provided as an input to the optimization model implemented in the advanced planning tool of WP 3.

In parallel to the development of the new planning tool, **Pan-European scenarios** are created in WP 4. These Pan-EU scenarios are elaborated for the years 2030-2040-2050 based on well-established EU and national "visions" as well as on the ENTSO-E's Ten-Year Network Development Plan (TYNDP) [5], European targets for the upcoming decades, national regulations and other relevant aspects such as climate targets. The developed scenarios consider different restrictions concerning primary energy resources, e.g. coal, gas or nuclear fuel, due to sociopolitical and economic aspects. While Pan-EU scenarios study potential key indicators affecting system planning and operation, the development of a fully pan-EU optimization of grid expansion case would require extensive simplifications due to computational restrictions, which are not

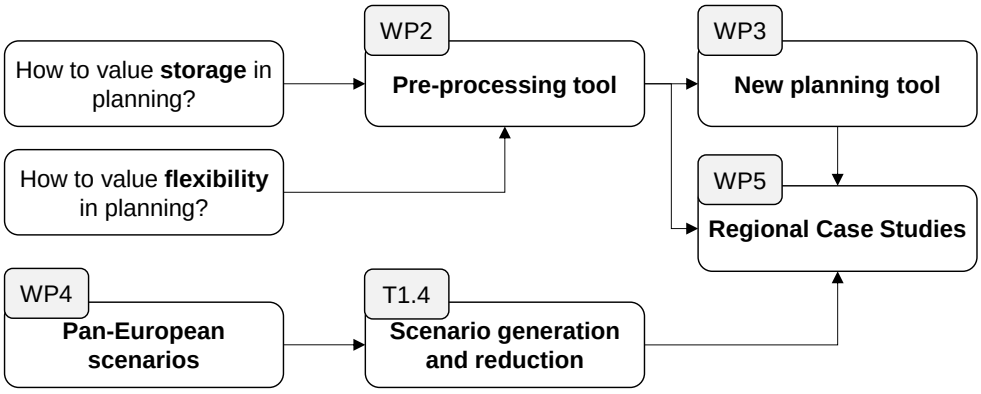

Figure 1-1: FlexPlan project conceptual block diagram

suitable for detailed network planning studies.

Thus, starting from the results of the Pan-EU macro-scenarios, six **Regional Cases** are developed in WP 5 in order to carry out detailed planning studies for the main European macro-zones considering the transmission and distribution grid infrastructure in detail. The preceding WP's interim results are combined in the regional cases to obtain the optimal grid expansion plan for each considered region. To do so, one hand, the macro-scenarios in terms of installed generation capacities, power demand and expected interzonal flows calculated by WP 4 Pan-EU scenarios are considered as input parameters for the planning tool. On the other hand, the pre-processor tool's results in terms of flexibility characteristics, potential locations of storage devices and potential transmission expansion candidates are considered as optimization variables. Additionally, the new planning tool of WP 3 that is implemented based on the specifications elaborated in WP 1 is utilized to carry out the regional case studies, which includes advanced modelling techniques, in particular:

- Multi criteria optimization
- Probabilistic security analysis
- Consideration of environmental aspects
- Monte-Carlo scenario generation and reduction

Task 1.4's scenario reduction approach will be directly integrated in the new advanced planning tool that is implemented in WP 3 as an optional module. Thus, potential expansion candidates are identified with WP 2's pre-processor tool based on the initial power flow results of WP 5's regional cases considering the

WP 4's reference macro-scenarios which are defined and time series data for a specific year without detailed consideration of stochastic inputs. As such, the methodology presented within this task provides the necessary link between Pan-EU macro-scenarios and the time series input required by the planning tool, including the stochastic representation of renewable generation and demand.

Objectives of this document

This document is the final deliverable of Task 1.4: *Scenario Generation and Reduction Approach*. The document:

- describes the approach to generate operational scenarios for wind, solar and hydro generation as well as load time series based on pan-EU scenarios of WP 4,
- describes the approach to identify a sub-set of representative operational scenarios by clustering techniques and thereby reduce complexity and calculation time of the planning problem,
- discusses the modelling choices that have been taken against the background of long-term pan-EU grid expansion studies,
- discusses state of the art techniques to model uncertainties in long-term grid expansion studies,
- presents a literature review highlighting alternative methods to the proposed method,
- and demonstrates the applicability of the method using a small-scale test case.

2. Definitions and applied Data

In this section, individual essential terms and parameters are defined and the specific input and output data of the methodology is described.

2.1. Sets, indices and data model

Variables	Symbol	Indices
Set of planning horizons	S_y	$y \in S_y$
Set of periods in the planning horizon	S_t	$t \in S_t$
Set of nodes	S_n	$n \in S_n$
Set of generators	S_g	$g \in S_g$
Set of flexible demand elements	S_u	$u \in S_u$
Set of time steps	S_t	$t \in S_t$
Set of specific time periods	$S_{\Delta t}$	$\Delta t \in S_{\Delta t}$
Set of macro scenarios	S_m	$m \in S_m$
Set of historic observations	S_s	$s \in S_s$
Set of NUTS-2 regions	S_r	$r \in S_r$
Set of countries	S_c	$c \in S_c$

Variables	Symbol	Cardinality	Unit	Comment
Specific period of time	Δt		Hours	
Typical duration until a unit is back in operation	Δt^{OUT}		Hours	
Change in electricity demand in country *c* in hour *t* due to heating	$\Delta P_{c,t}^{\mathrm{Load,Heat}}$		MW	
Change in electricity demand in country *c* in hour *t* due to heating	$\Delta P_{c,t}^{\mathrm{Load,Cool}}$		MW	
Positive absolute difference in ambient temperature of a country *c* in hour *t* in operational scenario *s* and reference scenario s_{Ref}	$\Delta \vartheta_{c,t,s,s_{\mathrm{Ref}}}^{\mathrm{Pos}}$		°C	

Negative absolute difference in ambient temperature of a country *c* in hour *t* in operational scenario *s* and reference scenario s_{Ref}	$\Delta\vartheta^{\mathrm{Neg}}_{c,ts,s_{\mathrm{Ref}}}$		°C	
Capacity factor of a technology in a specific period of time Δt	$CF^{\mathrm{Tech}}_{\Delta t}$		p.u.	
Capacity factor of a technology in a specific NUTS-2-region *r* in hour *t* in an operational scenario *s*	$CF^{\mathrm{Tech}}_{r,t,s}$		p.u.	
Hydro capacity factor in a specific country *c* in hour *t* in an operational scenario *s*	$CF^{\mathrm{HY,RoR}}_{c,t,s}$		p.u.	
Number of day in year	d		-	
Active energy produced in a specific period of time Δ*t*	$E^{\mathrm{Gen}}_{\Delta t}$		MWh	
Linear load-temperature sensitivity factor for heating in country *c*	LF^{Heat}_{c}		MWh/°C	
Linear load-temperature sensitivity factor for cooling in country *c*	LF^{Cool}_{c}		MWh/°C	
Failure probability of thermal power plant	p^{Fail}		%	
Total active power generation in NUTS-2-region *r* in hour *t* for macro-scenario *m* and operational scenario *s*	$P^{\mathrm{Gen}}_{r,t,m,s}$		MW	
Installed run of river hydro generation capacity at node *n* in	$P^{\mathrm{HY,RoR,Inst}}_{n,m}$		MW	

macro-scenario *m*				
Active hydro power generation at node *n* in hour *t* for macro-scenario *m* and operational scenario *s*	$P^{\mathrm{HY,Gen}}_{n,t,m,s}$		MW	
Installed generation capacity at node *n* in macro-scenario *m*	$P^{\mathrm{Inst}}_{n,m}$		MW	
Total installed generation capacity in NUTS-2-region *r* in macro-scenario *m*	$P^{\mathrm{Inst}}_{r,m}$		MW	
Total active power demand at node *n* in hour *t* for macro-scenario *m* and operational scenario *s*	$P^{\mathrm{Load}}_{n,t,m,s}$		MW	
Maximum active power generation	$P^{max}_{g,t,y}$	$\forall g \in S_g, \forall t \in S_t, \forall y \in S_y$	MW	
Total active power demand at node *n* in hour *t* for macro-scenario *m* and reference operational scenario s_{Ref}	$P^{\mathrm{Load,Ref}}_{n,t,m,s_{\mathrm{Ref}}}$		MW	
Active power reference	$P^{ref}_{g,t,y}$	$\forall g \in S_g, \forall t \in S_t, \forall y \in S_y$	MW	For renewable generators (PV, Wind, Hydro), the active power reference equals the maximum active power generation.
Reference demand	$P^{ref}_{u,t,y}$	$\forall u \in S_u, \forall t \in S_t, \forall y \in S_y$	MW	
Average proportion of time in which a unit is out of order	r^{OUT}		%	
Random number between 0 and 1	x		%	

2.2. Geographic scope

In the FlexPlan project, **six regional cases** (RC 1 – RC 6) are developed, which will serve for both validating the practical application of the advanced planning tool, and also to create results which are necessary to analyze the order of magnitude of the expected benefits of flexibility for the system in the medium-long term horizon. The six regional cases depicted in Figure 2-1 cover the main macro-European zones, having each typical generation and network characteristics: Iberian Peninsula, France and Benelux,

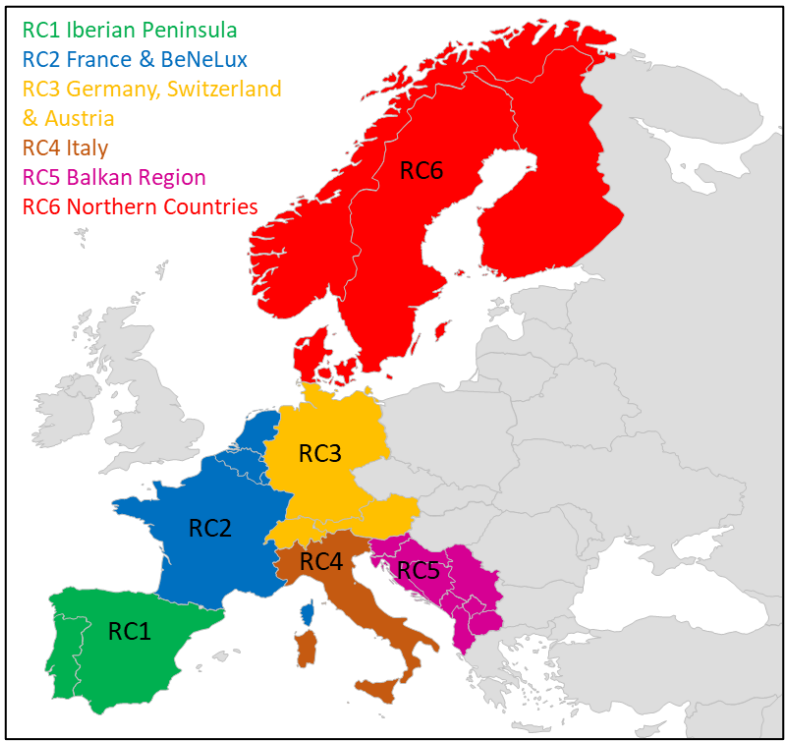

Figure 2-1: FlexPlan's six regional cases

Germany, Switzerland and Austria, Italy, Balkan region and Northern Europe.

The fundamental input for the regional cases is provided by TU Dortmund's *MILES* simulation framework, for smaller sub-regions in each country. A more detailed description of MILES can be found in the appendix of FlexPlan's **D4.1 Pan-European scenario data** [4].

The regionalization module of *MILES* calculates installed capacities at zonal level as well as time series for Renewable Energy infeed for 34 countries in Europe, starting from National level installed capacities defined in the Pan-EU scenarios. To distribute installed capacities of RES, *MILES* applies a top down approach. The National territory of each country is divided into a number of sub-regions. Subsequently, various statistical parameters for each region are analyzed carefully to generate specific regionalization factors. The considered statistical parameters include socio-structural data, land use, location of existing plants and climate characteristics. In a second step, *MILES* calculates feed-in time series for each region based on historical weather data and the assigned installed capacities. The weather data processed in *MILES* is taken from the regional model *COSMO-EU* of Germany's National Meteorological Service [6]. A historical load profile is broken down by means of regionalization factors and scaled to the targeted annual consumption, to generate spatially

disaggregated time-series of the electrical load in each sub-region. For the electrical load, a distinction must be made between the load of households, the service and industry sectors. Based on available data, either the number of households or the population of every region is used as main parameters to distribute residential demand. The regional distribution of the electrical load of the service sector is described by several parameters without considering any weighting parameters. The gross domestic product and population density are major indicators for the electrical energy demand. Other main parameters are the area of commercial buildings and related open space, as well as the working population of each sub-region. Table 2-1 provides the considered number of sub-regions for each country considered in the FlexPlan regional cases.

Table 2-1: Number of considered sub-regions per country per regional case

Regional Case	Country	sub-regions [#]
1	Spain	599
	Portugal	404
2	France	766
	Netherlands	37
	Luxembourg	11
	Belgium	46
3	Denmark	162
	Norway	168
	Sweden	175
	Finland	70
4	Italy	728
5	Serbia	79
	Macedonia	103
	Albania	165
	Montenegro	67
	Bosnia Herzegovina	240
	Croatia	233
	Slovenia	174
6	Germany	732
	Austria	70
	Switzerland	126

To ensure maximum consistency between the macro-assumptions taken in the pan-EU scenarios of MILES and the detailed nodal operational states that will be considered in the regional cases, detailed information on the sub-regional level is used as inputs for the scenario generation and reduction described in the present report. The detailed output of *MILES* provides the installed generation capacity for ***each node of the transmission grid provided by ENTSO-E as an individual sub-region the pan-EU results***. A full time-series of 8,760 hours is provided for each macro-scenario. This feature of *MILES* helps us to avoid extensive zonal to nodal data transformations in the following tasks and work packages. The spatial distribution of the considered transmission grid nodes in Europe is visualized in Figure 2-2.

However, the high level of detail in this tasks' input data (***approximately 5,100 locations in regional cases*** and further locations in Eastern Europe) requires an adequate scenario generation and reduction approach considering generation and demand stochasticity on a similar spatial and temporal resolution.

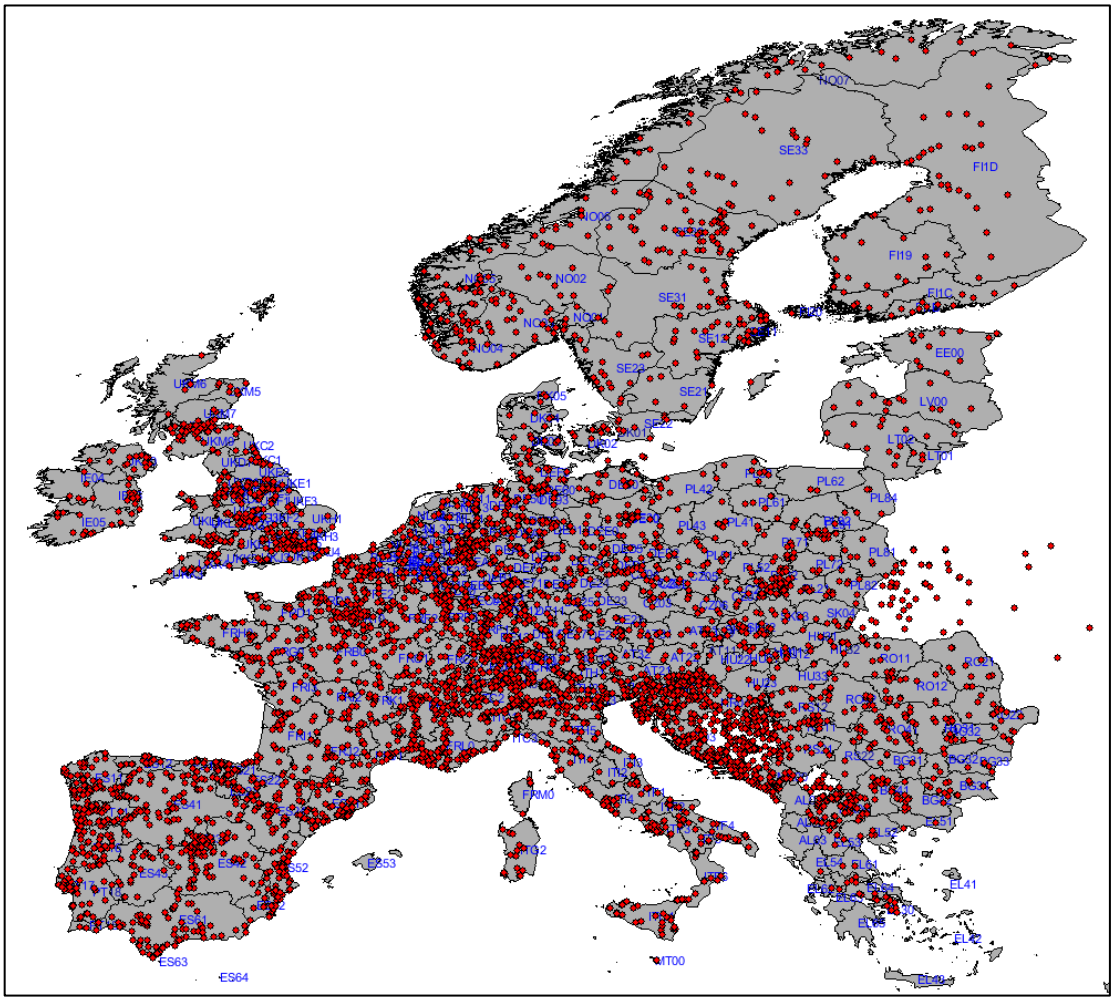

Figure 2-2: Transmission grid nodes in Europe considered as sub-region in MILES

The *Nomenclature of Territorial Units for Statistics* (NUTS) forming, the so called "***NUTS-Regions" has been identified as a suitable geographic reference system*** to model spatial and temporal uncertainties in this task, as there is sufficient data publicly available that can be integrated in the developed scenario generation and reduction methods. The geocode standard for the NUTS regions, is developed and regulated by the European Union, therefore it only covers the member states of the EU in detail.

For each EU member country there are in total four NUTS levels that are established by Eurostat in agreement with each member state. The current NUTS classification (2018) covers 28 regions at NUTS-0 level (EU Member States), 104 regions at NUTS-1 level, 281 regions at NUTS-2 and 1,348 regions at NUTS-3 level. In most cases, the classification is identical with the administrative structure of the Member States. However, not all NUTS-regions correspond to administrative divisions. The four hierarchical levels are:

- The population of all European **NUTS-1** regions is between 3 and 7 million inhabitants. However, there are exceptions to this rule, particularly in smaller EU states (e.g. Luxemburg, Cyprus, Estonia), where the **NUTS-0** region (Member State) often also represents the NUTS-1 and

sometimes even NUTS-2 level. In Germany for example the NUTS-1 regions correspond to the federal States (Bundesländer).

- **NUTS-2** regions usually have between 800,000 and 3 million inhabitants. In the United Kingdom for instance, this level represents mainly the level of the counties.
- **NUTS-3** regions generally have a population of 150,000 to 800,000 inhabitants. In France, this is the level of the departments, whereas in Belgium it is the level corresponding to the Arrondissements.

The NUTS-2-Level offers a good compromise in terms of spatial accuracy, data availability and data handling complexity. The regions of the NUTS-2-Level (2013) are visualized in Figure 2-3.

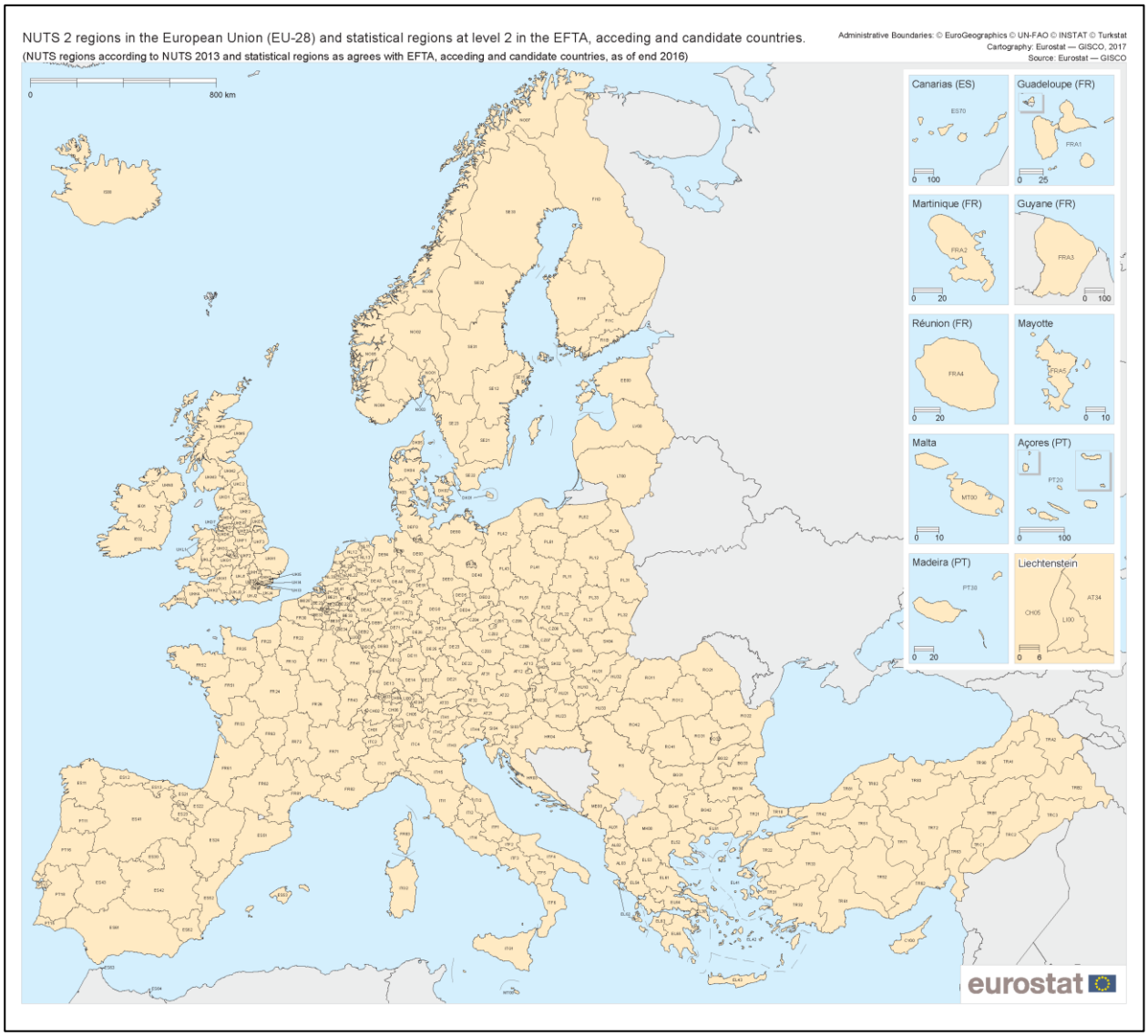

Figure 2-3: NUTS-Regions (Level 2) in Europe (2013) used as geographic reference system to model uncertainties

3. State of the Art in Scenario Generation and Reduction

The aim of transmission and distribution network expansion planning is to find a set of reinforcement and expansion measures forming a robust grid configuration that serves the load reliably at minimum cost for a large variety of grid use cases.

Since the power system's degree of penetration with non-dispatchable renewable energy sources increases more and more, the total bandwidth of possible grid use cases has grown significantly. As a consequence, ***traditional planning approaches considering only a few selected hours as inputs, e.g. the hours with highest and lowest load, are no longer suitable*** to assess future energy system's adequacy and reliability. Thus, advanced methods have to be applied in power system planning to incorporate the extended bandwidth of possible grid use cases in the future.

As the share of non-dispatchable RES in the system rises, power system operation, control and planning have to cope with increasing amounts of uncertainties. ***Uncertainties in the field of power systems that are considered in FlexPlan*** and its scenario generation approach respectively***:***

- unplanned ***outages*** of thermal units,
- intermitting ***generation of renewable energy sources***,
- weather / temperature sensitive electrical ***load*** and
- amount of available ***hydropower generation***;

These uncertainties, referred to as ***stochastic inputs of the planning problem***, are significantly impacting power system operation as well as power system reliability already today and will continue to do so in the future. Thus, advanced transmission and distribution ***grid expansion planning methods have to incorporate these stochastic inputs*** in an adequate way to be able to deliver a robust solution. A solution is considered as robust, if it the determined grid infrastructure is able to supply the demand in all conceivable credible situations. To model such kind of extraordinary conditions, adequate stochastic inputs have to be chosen incorporating credible conditions, such as droughts, cold spells, windstorms, high-pressure weather conditions and random combinations of the aforementioned conditions resulting in critical events by so called **scenarios**.

Optimization methods and/or heuristics are often used to determine the subset of cost-optimal grid reinforcement and expansion measures and their optimal positions in the existing infrastructure. The existing methods differ mainly in the applied solution algorithms as well as in the considered time period. In the field of mathematical programming optimization, decomposition approaches according to Benders (cf. [7] [8] [9] [10]) are often applied, whereas evolutionary algorithms [11] [12] are widely used in the field of heuristic solution algorithms. ***The grid expansion problem often only considers one specific target year (static planning)*** [11] [12]. Increasingly, however, the time-coupled consideration of several consecutive planning periods is getting in the focus of attention, whereby so-called dynamic approaches (cf. [10]) are gaining in popularity. However, the ***complexity of the problem increases significantly, especially when several periods (dynamic planning) are considered*** [13]. In existing approaches to solve the grid expansion problem, uncertainties in the input variables are often represented only insufficiently or in a very simplified form in order to keep the problem complexity at an acceptable level.

A **scenario** typically includes forecasts of the installed capacity per generation technology and the load in the target year based on a so-called storyline, which qualitatively describes the political and legal framework. In order ***to depict the totality of all possible future scenarios, typically three to four extreme or marginal scenarios are defined***, which form the largest possible scenario funnel and thus cover the broadest possible spectrum of uncertainties. Based on the identified extreme/border scenarios, detailed investigations are typically carried out on an hourly basis for one year (8,760 hours), considering historical climatic and hydrological conditions with regard to the necessary grid expansion requirements. Recently, ***a large number of historical weather years*** for modelling the supply-dependent generation from wind and sun have been taken into account ***in electricity market-based analyses*** of system's adequacy [14], whereas ***even today only one typical weather year is taken into account in detailed studies of grid expansion requirements*** [15].

In addition to the scenarios published by transmission/distribution system operators for the development of the national and international energy supply system, there are a large number of other scenarios [16], which have been developed by research institutes and other institutions. ***It should be noted that the forecasts contained therein for the future development of the system sometimes differ significantly from the established extreme/border scenarios. As such, we can conclude that the frequently practiced use of only three to four extreme or boundary scenarios in network planning is not sufficient to cover the entire range of uncertainties in the input data.***

3.1. Consideration of uncertainties in power system planning

The model-based determination of the future grid load under uncertainty is the subject of numerous studies and research projects. A fundamental distinction must be made between the **operational and planning perspectives**. In the context of transmission system operation, uncertainties in the short term have to be considered as exactly as possible in order to adequately consider risks of critical grid situations in advance. On the other hand, uncertainties in long-term grid planning must be considered in order to reduce risks associated with extensive investments in the grid infrastructure. In the field of long-term grid planning in particular the risk of sunk cost and stranded investments are present. The main focus of this research project is on long term grid planning. As such, the modeling of stochastic input variables is more focused on **long-term uncertainties** (climatic and hydrological conditions) and less on short-term uncertainties, e.g. wind and load forecast errors or power line and power plant outages.

There are several different approaches to model the long-term uncertainties and their interactions with each other in the context of grid planning. On the one hand, uncertainties in the input parameters can be explicitly represented by their stochastic properties. Therefore, it is necessary to define a **probability distribution function** for all uncertain input parameters. On the other hand, the consideration of **all combinatorial possibilities** of the uncertain input parameters (at least in theory) represents a further option. Thereby, all theoretically possible combinations of uncertain parameters' realizations are examined. This procedure is equivalent to a brute force method or an **exhaustive search** approach. Furthermore, it is possible to represent the uncertainties in the overall problem only by a limited number of deterministic **extreme scenarios**, assuming these extreme conditions include all possible realizations. Setting up a (multivariate) distribution function is often very complex and not applicable to all problems. However, multivariate distribution functions are able to model all dependencies between the uncertainties in a compact way. In contrast, a limited number of extreme scenarios is easily set-up, but they are obviously not able to

model the full range of the uncertainties. Furthermore, the consideration of all combinatorial possibilities is often not possible due to the large number of uncertainties and their individual characteristics, such that a problem reduction is carried out by selecting presumably extreme scenarios.

For the consideration of uncertain input variables, such as intermittent renewable energy sources, variable hydro generation or temperature-dependent loads, stochastic modelling techniques were developed. In contrast to deterministic methods, all (or a selected sub-set) of input and output variables are treated as random variables. By assuming that uncertainties or variations in the input variables can be estimated or measured, the probability density functions of the input variables are known [17]. In case uncertainties in the respective input variables cannot be estimated or measured, assumptions have to be made. The aim of the probabilistic methods is to determine probability density functions of the output variables to measure their respective variability. Meanwhile, many different probabilistic procedures exist, which can be divided in two main groups: Monte-Carlo simulations and analytical procedures.

If the uncertainties were described in terms of scenarios, only the Monte-Carlo simulation can be applied or a transformation of the scenarios into a multivariate distribution function is necessary. Most of the existing probabilistic approaches require known, continuous probability density functions comparable to a normal distribution as input variables. If the input variables are only partially known or discretely distributed, most of the probabilistic methods become inaccurate. Thus, they are no longer applicable.

The **Monte-Carlo simulation** is often used in the analysis of energy systems. It is regarded as the simplest solution method for probabilistic problem formulations and is able to deliver precise results [17]. At the beginning of each iteration, random values are sampled from the input variables, subsequently a deterministic problem is generated and solved. With the Monte-Carlo simulation, there are no limitations regarding the description of input variables. On the one hand it is possible to use probability density functions and on the other hand time series data can be used with Monte-Carlo approaches as well. Thus, Monte-Carlo approaches are very flexible with regard to the inputs and therefore can be perfectly adapted to specific problem types. However, a large number of iterations is required to achieve convergence [18]. Consequently, the computational effort is very large [17] and increases with growing number of input and output variables [19].

In contrast to the Monte-Carlo simulation, **analytical methods** are computationally more effective. They determine the relations between the input and output variables by simplifying the probabilistic problem by mathematical assumptions. Analytical methods include convolution, Fast Fourier Transformation and fuzzy logic [18]. The convolution method is a common widely used method. With this approach all possible combinations are considered, but it has two major disadvantages. On the one hand, a linearization of nonlinear relations is necessary. The second disadvantage is the assumption of complete statistical independence of the input data. It is therefore difficult to apply to real life uncertainties.

Furthermore, the incorporation of uncertainties in probabilistic models can be differentiated in analytical approaches that completely rely on **probability density functions** as inputs whereas some Monte-Carlo approaches work with huge amounts of **time series data** coming from a significantly large data pool of (historical) observations or measurements. Especially, modeling uncertainties in power generation of intermittent renewable energy sources as well as in the electrical load can be used to illustrate these two fundamentally different approaches.

To model the power generation of wind turbines, historic weather conditions are used in some cases, i.e. a **time series** of historic wind speed near the plants' locations. The recording of one historic year consisting of 8,760 hours is used subsequently to determine the dispatch as well as the grid utilization [20]. On the one hand, this procedure requires a high computational effort compared to an analytical approach using an adequate distribution function to generate the same number of samples. This is especially the case, if probability distributions are treated analytically rather than numerically in the subsequent analysis. ***On the other hand, it is known that the consideration of only one historic year (resulting in 8,760 samples) is not representative for future years, as climatic conditions (especially wind speed) can vary considerably from one year to another.*** However, the consideration of the full time series is advantageous, as the intertemporal structure is not lost, against the background of the resulting gradients in wind power generation from one hour to another. In [18] and [21] the wind speed is represented by a Weibull distribution. The dependencies are modeled by the Pearson coefficient and a small number of wind turbines is considered. Some research works use copulas, which allows a flexible modeling of dependencies. The concepts of Copulas was introduced in 1959 by Abe Sklar and is mainly used in the fields of actuarial and financial mathematics. However, more recent works [22] apply the concept of copulas as well in the field of energy system modelling. A copula is a function that models the dependency structure between different random variables. Here, the marginal distributions and the dependency structure are modeled independently. The Copula is a flexible tool to consider both linear and nonlinear dependencies [23]. In the previous work, only small Copula models are built, which consider a small number of dimensions. Usually only two or three variables are connected by a copula [24] [25] [26] [27]. A much higher number of dimensions is necessary by considering realistic large-scale infrastructures as the European transmission grid. Each uncertain variable at a specific location or node goes along with an additional dimension, resulting in a complex high-dimensional copulas. In [28], uncertainties of power plant and line failure as well as in the load forecast are illustrated by scenarios. Uncertainties in the load are often represented by Gaussian distributions, e.g. in [29] [30] [31]. In [30] the load is additionally modelled with a gamma distribution. Furthermore, many different uncertainties are represented by “extreme scenarios”, e.g. in [32]. ***However, the generation of these multiple scenarios (in terms of different generation and load time series) does not follow a closed methodology. Thus, there is a need for further research on how to incorporate uncertainties in time-series based large-scale transmission expansion planning by a closed scenario generation and reduction method.***

In **summary**, it can be concluded that analytical methods work comparably quickly, but have only limited flexibility with regard to problem formulation. The use of analytical methods and the associated determination of probability density functions is often accompanied by unavoidable simplifying assumptions. In contrast, the iterative solution algorithm by means of Monte-Carlo simulation takes comparatively long time. However, iterative problem solving offers a higher degree of flexibility with respect to the problem formulation itself, whereby individual input variables can be treated separately with respect to their individual uncertainties. In addition, the complexity of implementing an analytical approach increases disproportionately with increasing size of the considered system compared to a Monte-Carlo simulation, because all uncertain input data - with their respective stochastic properties - have to be modelled explicitly**,** ***including*** their correlations. Furthermore, the consideration of intertemporal properties, such as hydro reservoirs’ levels of storage, in analytical procedures is only possible with additional considerable effort (sequential approaches, e.g. [33]). By deterministic consideration of the problem per iteration in a Monte-Carlo simulation, it is comparatively easy to model intertemporal properties that couple several time steps.

For the reasons mentioned above, the modeling of the stochastic input variables by the consideration of a broad variety of historical weather conditions a Monte-Carlo approach is therefore chosen within the framework of this project.

3.2. Scenario generation techniques (time series based)

As the FlexPlan projects aims at explicitly incorporating storage and demand flexibility in the planning process, the consideration of consecutive time steps, i.e. time series data, is essential. ***Hence, hourly time series are generated for the afore-mentioned stochastic input parameters.*** Spatial and temporal correlations need to be incorporated correctly when generating time series for intermittent renewable energy sources and the electrical load as well. ***Thus, each time series' individual seasonality and trend has to be identified as well as correlation between the time series.***

Basically, there are two approaches to tackle the problem of spatial and temporal correlations in the scenario or especially the time series generation process. One the one hand **statistical methods** can be used to analyze historical data to identify the intrinsic characteristics of each time series namely wind, solar and hydro power generation and load. **Numerical weather models** can be applied to model the feed-in of intermitting renewable energy sources bottom-up on the other hand.

3.2.1. Statistical methods

Statistical methods make use of historical measurements to train a **statistical model** (e.g. neural network or comparable machine learning algorithms) to transform the data. However, the application of statistical methods in the field of power system planning has several problems, as it is heavily depends on high quality training data to learn the intrinsic characteristics of the respective generation and load time series as well as their correlations. On a national level, historic data of total variable Renewable Energy Sources (vRES) feed-in and total load might be publicly available for several years, but spatially disaggregated feed-in and load time series (e.g. per sub-station) are typically not publicly available, especially not for several years. As a consequence, statistical methods could be used to identify and learn the intrinsic characteristics on an aggregated level, but due to the lack of adequate training data (spatially disaggregated wind, solar and hydro power generation) it is not possible to identify spatial correlations. Especially, by analyzing the future energy system it is not possible to have detailed past production data of vRES, as they will be built in the future. Thus, the application of statistical methods is not applicable for prospective system studies. Furthermore, individual production data of vRES generators, e.g. wind farms, is typically not publicly available, as it is treated as confidential in most cases. Hence, alternative methods and techniques have to be applied to generate consistent scenarios in terms of detailed spatial and temporal vRES feed-in and load time series of future the future energy system. ***As a consequence, the transformation of numerical weather data into electrical power outputs has become an independent field of research: energy meteorology; delivering suitable techniques to generate temporal and spatial correlated vRES time series. Thus, the FlexPlan scenario generation approach makes use of numerical weather model's data as well as physical models to create consistent scenarios.***

For the sake of completeness, a detailed review of statistical methods for scenario and time series generation can be found in [67] [68].

3.2.2. Energy meteorology and physical models

Due to the predominantly fossil-based energy system, which uses storable and transportable energy sources (coal, lignite, oil, gas and nuclear), meteorological topics were for a long time only of minor importance in energy system analysis. This situation already changed fundamentally at the turn of the millennium with the beginning of the increasing use of renewable energy sources. This led to the foundation of the still young field of energy meteorology. In particular, the renewable primary energy sources wind and solar PV are planned as fundamental components of energy generation in the future energy system. However, these cannot be stored and are volatile in their supply. Due to this increasing weather-dependent energy production, comprehensive information about the spatial and temporal availability of these energy sources is needed to simulate the energy system [34].

In addition, a detailed understanding of the influence of meteorological parameters is necessary in the planning of the future energy system. Power generation of intermittent renewable energy sources is mainly determined by their location and the corresponding local climatic conditions. To capture and describe these interactions between meteorological and energy processes, existing meteorological models are used. These numerical simulation models map the physical effects of the atmosphere based on the basic equations for momentum, mass and energy conservation as well as further balance equations for cloud water, rain water and precipitation particles. One result is a three-dimensional wind field, which considers the influence of orography in layers close to the ground. For example, the influence of forests or settlements is mapped as well as the increase of wind speed over crests and mountain ridges. Thus, by the application of numerical weather models inaccurate extrapolations of data from near-ground measuring stations to determine the wind speed at different heights, e.g. hub height of a wind power plant, can be avoided, as this process typically goes along with increasing uncertainty depending on the height [35] [36] [37].

There is a multitude of **weather models** in the literature. The calculation results of some of these models are publicly available. An overview of selected models can be found in the following Table 3-1.

Table 3-1: Overview of publicly available weather models (reanalysis datasets)

Institution	Model	Coverage		Time resolution	Spatial resolution	Model heights [m]
DWD	COSMO-DE	2005	present	1 h	2.8 km x 2.8 km	10, 36, 73, 122, 184
DWD	COSMO-EU	2006	2016	1 h	7 km x 7 km	10, 35, 69, 116, 179
DWD	ICON	2017	present	1 h	7 km x 7 km	10, 43, 99, 174
ECMWF	ERA40	1957	2002	6 h	80 km x 80 km	10, 60
ECMWF	ERA-Interim	1979	present	6 h	80 km x 80 km	10, 60
ECMWF	ERA-20C	1900	2010	3 h	80 km x 80 km	10, 91, 100
ECMWF	ERA5	1979	present	1 h	30 km x 30 km	10, 100, 137
JMA	JRA-25	1979	2004	6 h	60 km x 60 km	10, 40
JMA	JRA-55	1958	present	6 h	60 km x 60 km	10, 60
NASA	MERRA-2	1980	present	1 h	55 km x 70 km	2, 10, 50, 72
NCEP	R2	1979	2012	6 h	278 km x 278 km	10, 28
NCEP	CFSR	1979	2010	1 h	35 km x 35 km	6, 10
NCEP	CFSv2	2011	present	1 h	38 km x 38 km	6, 10
NOAA	20CRv2	1871	2011	6 h	222 km x 222 km	10, 28

The models differ in the available historical time range, the temporal and geographical resolution and the model heights. All listed models are constructed according to similar principles, but differ in the scope of the systems of equations, the type and number of parameterization or the initial and boundary conditions. For the modelling of the feed-in of the wind energy plants, however, the regional resolution and the simulated height layers of the models are particularly decisive. The weather models of the organizations National Aeronautics and Space Administration (NASA), Global Modeling and Assimilation Office (GMAO), National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Prediction (NCEP), Japan Meteorological Agency (JMA) and European Centre for Medium-Range Weather Forecasts (ECMWF) simulate the entire earth's atmosphere and are therefore called global models. Due to the large amount of data and the required computing power, only models with a less detailed spatial resolution are available. The models of the German Weather Service (DWD) are local models and focus on Germany or Central Europe. The models and the simulated data are provided by government institutions and are freely accessible to the public.

Some of these global models are already being used as a database to simulate the supply of renewable energies [38] [39] [40] [41] [42]. ***An evaluation of the models ERA5, MERRA-2 and COSMO-REA6*** *(successor of COSMO-EU)* ***with regard to their applicability in power system simulations can be found in*** [43] where a detailed case study for wind production in France was carried out. Furthermore, a detailed comparison of the applicability of the aforementioned weather models to simulated intermittent renewable energy feed-in can be found in [44].

3.2.3. Transforming numerical weather data in electrical power outputs

Power system studies such as transmission expansion planning need power generation of all generators as an input. Thus, the feed-in of non-dispatchable renewable energy sources like wind and solar has to be modeled in the scenario generation approach. Therefore, the meteorological information of numerical weather models has to be converted in electrical power outputs.

In contrast to the statistical models, **physical models** make use of turbine power curve functions to transform the wind speed at hub-height to electrical power generation. The application of physical models requires at least some fundamental information on the technological parameters. On the one hand, the model or type of the wind turbine is used to identify an adequate matching power curve. On the other hand, the hub-height of the plant has to be known to extrapolate the wind-speed at this level. ***One major drawback of physical models is the potential bias included in the underlying wind-speed data. As a result, physical models tend to over or under-estimate the power outputs. Thus, physical models are typically combined with bias correction approaches***, e.g. [39] by comparing modelled data with historic data to derive correction factors [43]. The need for bias correction techniques results from the fact that weather reanalysis data is calculated with computer models that are less than perfect and include systematic errors (biases) due to errors in the underlying numerical weather model [39].

Physical models - Wind power

Currently, reanalysis data sets for the simulation of feed-in time series of wind turbines are increasingly being investigated. First investigations of wind power generation using reanalysis data sets were carried out for selected sites in Hungary [45] and Northern Ireland [46]. In recent years, studies with larger areas of consideration, typically national studies, have been carried out, e.g. for UK [47] [48] [49] [50], Denmark [51]

and Sweden [52], in order to validate the simulated time series by comparison with recorded data. It should be noted, however, that the previous studies only concentrated on comparatively small and geographically similar areas. Furthermore, these studies only validate the modelled data on the basis of wind speeds and not on the basis of generation data, thus neglecting the essential part of the model - the transformation of meteorological data into electrical generation data [39]. A comprehensive overview of current studies and the validation methods used in them can be found in [50]. In addition, there is an increasing number of publications [54] [55] [56] [57] [58] that deal intensively with the transformation of meteorological data into electrical generation time series, but do not deal with the validation of the results. Wind energy capacity factors throughout Europe are over- or underestimated by up to ±50%, as [39] noted.

Physical models – Solar power

As stated in [38] MERRA and MERRA-2 of NASA [59], ERA-Interim of the ECMWF [60] and JRA-55 of the Japanese Meteorological Agency [61] are frequently used for global reanalysis of the latest renewable generation. There are numerous new publications on the use of reanalysis data for wind energy simulation (e.g. [47] [49] [50] [51]). There seem to be two main reasons why reanalysis data is not used often for modeling solar generation. Firstly, the installed capacity of photovoltaics (2012) reached a level of 100 GW worldwide much later than wind energy (2008) [62]. Secondly, satellite images are an alternative open data source to obtain data on solar radiation in high spatial and temporal resolution. In addition, it is also possible that the modelling of solar feed-in is regarded as less challenging due to its typical recurring structure in the seasonal and diurnal course.

Nevertheless, there are some recent studies [55] [63] [64] [65] that use reanalysis data to simulate solar feed-in. However, these studies do not validate the weather data itself, as they accept it as it is. In [66], PV time series generated by MERRA are validated on the basis of two historical years of national aggregated PV production in Germany. In view of the fact that the suitability of reanalysis data for the simulation of PV feed-in for Europe-wide studies has not yet been conclusively clarified, since no validation against historical data with spatially and temporally high-resolution simulations for Europe has been carried out, [38] has made up for this. In [38] it was found that solar irradiance data of reanalysis models is basically suitable and adequate to be applied in energy system modelling. However, solar irradiance data of reanalysis is subject to fundamental biases and should be corrected, just as wind speed data [38].

Conclusion

As there is a multitude of weather models as well as temporal and spatially resolved production data derived from them publicly available, ***the scenario generation approach of this project makes use of*** *data coming from MERRA-2 and* ***physical models transforming numerical weather data in electrical outputs.***

3.3. Scenario reduction techniques

The objective of the network expansion planning tool is to find the optimal network expansion measures that will allow the network to operate reliably for a range of uncertain future conditions, spanning several decades (2030-2040-2050). These future conditions are characterized by several long-term visions, describing possible developments of the energy system as well as divergent European energy policies.

As stated above, uncertainties in the planning problem are introduced by the presence of renewable generation resources, temperature-dependent loads and hydro-condition dependent storage and production. With the robust approach in mind, the expanded network found as solution by the planning problem must be able to supply the demand in most, if not all, possible situations. As explained above, the approach chosen within FlexPlan is to provide a representative set of inputs, referred to as Monte Carlo years (MC years), to the planning tool to make sure that the solution is calculated based on a representative set of possible uncertain input scenarios, mainly characterized by the weather conditions.

However, due to the fact that the FlexPlan tool aims at covering a time horizon of multiple decades, including many technologies and spanning a large geographical area, solving this problem is a computationally demanding task. Therefore, the chosen set of MC-years must be limited as to allow for a computationally tractable planning problem. However, it is difficult to determine which scenarios are relevant a priori.

Before renewable generation became a substantial part of grid operations, it was often considered sufficient to evaluate new transmission lines only for the hour of the year with the highest load. Any transmission network that allowed for adequate operation during this hour was considered likely to operate at least as well during any other demand scenario. This “worst hour” approach to selecting a test scenario is still deceptively appealing, but in modern systems, it is not obvious which hour is most likely to cause issues that severely threaten the reliability and/or security of supply. Also, the worst hour may vary by region or change as new lines are added and reliability issues are resolved. The inclusion of storage (and other flexible technologies) into the transmission expansion plan, require representative time series as input scenarios to the planning problem to enable a correct modeling of the intertemporal constraints of the flexible assets.

In the literature on transmission expansion planning, a variety of strategies is described to reduce the overall space of possible network operating conditions to a smaller subset of single-hour operating conditions, often referred to as ‘time-slices’ [73]. These include the use of heuristics [77], Monte Carlo sampling, Latin hypercube sampling, forward and backward scenario selection [74], K-means clustering [72] [78] [75], robust optimization formulations, importance sampling [79], and various hybrids of the above methods [69]. However interesting, those techniques are focused on finding a limited number of ‘single-hour’ scenarios, and disregard the required time dimension of the scenarios for a planning problem including flexibility and/or storage. Moreover, it has been shown that for systems with a high penetration of RES, using a limited number of single-hour time-slices as input to a generation expansion planning problem leads to an underestimation of the variability of RES, hence to an underestimation of the value of flexible technologies [82].

Multiple ways to improve the modeling of the temporal dimension have been developed. Mainly in the field of generation expansion planning (e.g. LiMES-EU [71]) or energy system optimization models (e.g. TIMES [70]), scenario reduction is used to obtain a limited set of representative time series that reflect the variation of overall demand and generation over time. In the literature focusing on transmission network expansion planning, the topic of scenario generation and reduction is less discussed, as within these works the focus mainly lies with the mathematical formulation of the planning problem itself, rather than retrieving suitable scenario inputs. Although the transmission network itself is mostly ignored (or simplified) in the energy system optimization or generation expansion planning models, the scenario inputs required for these models are to a large extent similar to the inputs required for transmission network expansion planning. Moreover,

the conditions that define a representative set of input scenarios are largely similar for transmission network expansion planning as for the energy system optimization models.

Within the generation expansion or energy system optimization models, often a set of (well-chosen) representative days is used to represent an entire year. In many works, these representative days are obtained by using simple heuristics [83][84]. Generally, in most of these simple heuristic approaches, a number of periods with different load and/or meteorological conditions is selected in order to capture a variety of different events. For example, to select three representative days, one could select the day that contains the minimum demand level of the year, the day that contains the maximum demand level and the day that contains the largest demand spread in 24 hours, as used in [84].

A more advanced technique is to use clustering algorithms to find periods with similar load/generation patterns. With clustering-based scenario reduction techniques, the aim is to group the overall space of possible scenarios into a (smaller) number of clusters, based on a metric that characterizes the scenario. Then, one scenario from each cluster has to be selected to feed into the planning problem. Alternatively, a 'synthetic' scenario, formed as the 'average' of the scenario's present in the cluster, can be created as cluster representative to feed in to the planning problem. Additionally, to indicate the importance or probability of the scenario, these can be assigned a weight. This weight equals the size of the cluster (i.e. the number of periods that are present in the cluster)[1]. Clustering approaches may thus implicitly determine the weight assigned to every selected representative period, which allows to appropriately account for both common and rare events. This is the major advantage compared to the heuristic approaches. Many clustering algorithms exist, and are used to cluster input data for network expansion planning problems. For example, in [75] Ward's hierarchical clustering is used, and K-means clustering is used in [85] and [86].

A third method to select representative periods for planning problems are so-called probability distance-based scenario reduction techniques. With this approach, the selection procedure is directly based on evaluating the full set of representative periods using a pre-specified 'cost-function'. This cost function represents the 'distance' between representative sets. The idea is then to solve an optimization problem to find the reduced set of representative scenarios with the minimal distance with respect to the full original set. The distance or cost function includes the impact the scenario has on the final solution of the problem. This is a hard problem to solve, for which possible solution techniques described in literature make use of the fast forward and backward algorithm [87] [88] [89]. The drawback of these methods is that they are computationally heavy, since they involve solving an often complex optimization problem.

In [73] an assessment of different scenario reduction techniques used for generation expansion planning problems is made. One could argue that for transmission expansion planning, similar requirements exists on the selection of representative input scenarios. In [73], it is shown that on a limited test case the use of heuristics to determine a representative set performs worst, whereas the probability distance -based technique performs best, but at a high implementation and computation cost. The clustering-based technique performs sub-optimal, although much better than the heuristics-based method, and at a much lower computational cost. ***Therefore, in this project is opted to base the scenario reduction method on clustering.***

[1] The underlying assumption is taken that all scenarios in the original set of scenarios have the same probability weight.

4. Scenario Generation Approach

On the pan-European level a number of long-term visions, describing possible developments of the energy system and divergent European energy policies exist. Within the FlexPlan project, three visions for each target year based on diverse storylines, resulting in a total number of nine pan-EU scenarios (see D4.1 [4]) have been defined. Considering these visions as projections of the energy systems' development a **scenario funnel** is spanned, as depicted in Figure 4-1.

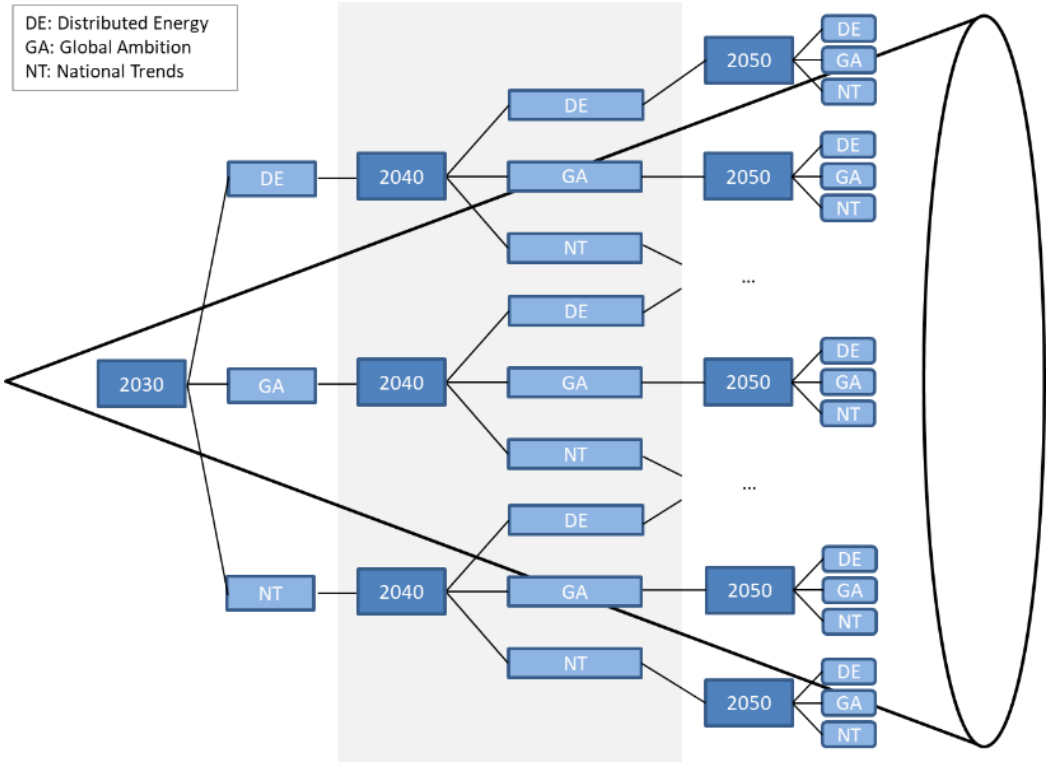

Figure 4-1: Scenario funnel formed by WP 4's Pan-EU long-term macro-scenarios

The Pan-EU scenarios developed within FlexPlan include macro-assumptions with regard to

- generation mix and demand per country,
- installed generation capacities and their technical properties, e.g. efficiencies,
- technical lifetime of power plants and expansion projects,
- fuel prices and availability of (synthetic) fuels,
- land usage and area potentials for renewable energy sources,
- new consumers and technologies impacting the total electricity demand.

It has to be noted that some macro uncertainties, especially in the field of installed capacities and technological progress, are already covered by the different visions in WP 4's Pan-EU scenarios. These macro uncertainties are referred to as **first stage uncertainties** that are out of the scope of the tasks of scenario generation and reduction approach, which is dedicated to generate operational scenarios including **second stage uncertainties** in terms of climatic and hydrological conditions.

To obtain operational scenarios, the Pan-EU scenarios on country level provided by *MILES* of TU Dortmund, are broken down to smaller geographic regions. Thus, national installed capacities per

generation and load category are spatially disaggregated to significantly smaller sub-regions in each country (see **D4.1** [4]). To this aim, within FlexPlan ***each node of the transmission grid model*** provided by ENTSO-E [95] is used ***as a sub-region within MILES***. As a consequence, the disaggregated scenario data per sub-region coming from WP 4 of FlexPlan can directly be used as an input for the scenario generation and reduction methods that are dedicated to the nodal level. Furthermore, ***the need for zonal to nodal data transformation processes is no longer necessary*** and beneficial for a ***straightforward integration of the scenario generation and reduction approach in FlexPlan's complete toolchain***, considering WP 2's pre-processor tool and WP 5's non-expanded optimal power flow simulations. The spatial disaggregation of the electrical load as well as renewable energy sources is only carried out once with *MILES*. Thus, the regionalization of WP 4's macro-assumptions itself is not considered as an uncertainty either.

The developed scenario generation approach within FlexPlan is depicted in Figure 4-2 and focuses on the **generation of operational scenarios in terms of hourly time series data** for renewable energy sources, hydro generation and load that are typically influenced by ***second stage uncertainties in terms of climatic and hydrological conditions***. As FlexPan's planning tool explicitly incorporates storage and demand flexibilities, the intrinsic characteristics of the generation and load time series over time have to be considered. Thus, a Monte-Carlo approach is applied to generate consecutive, correlated time series data based on historical climatic conditions. To generate time series of variable renewable energy sources, weather data for 40 historical years is used in combination with physical power plant models in two Time Series Generation (TSG) modules. Hydrological information of 36 historical years is processed to generate hydropower generation time series based on statistical models applied by [101].

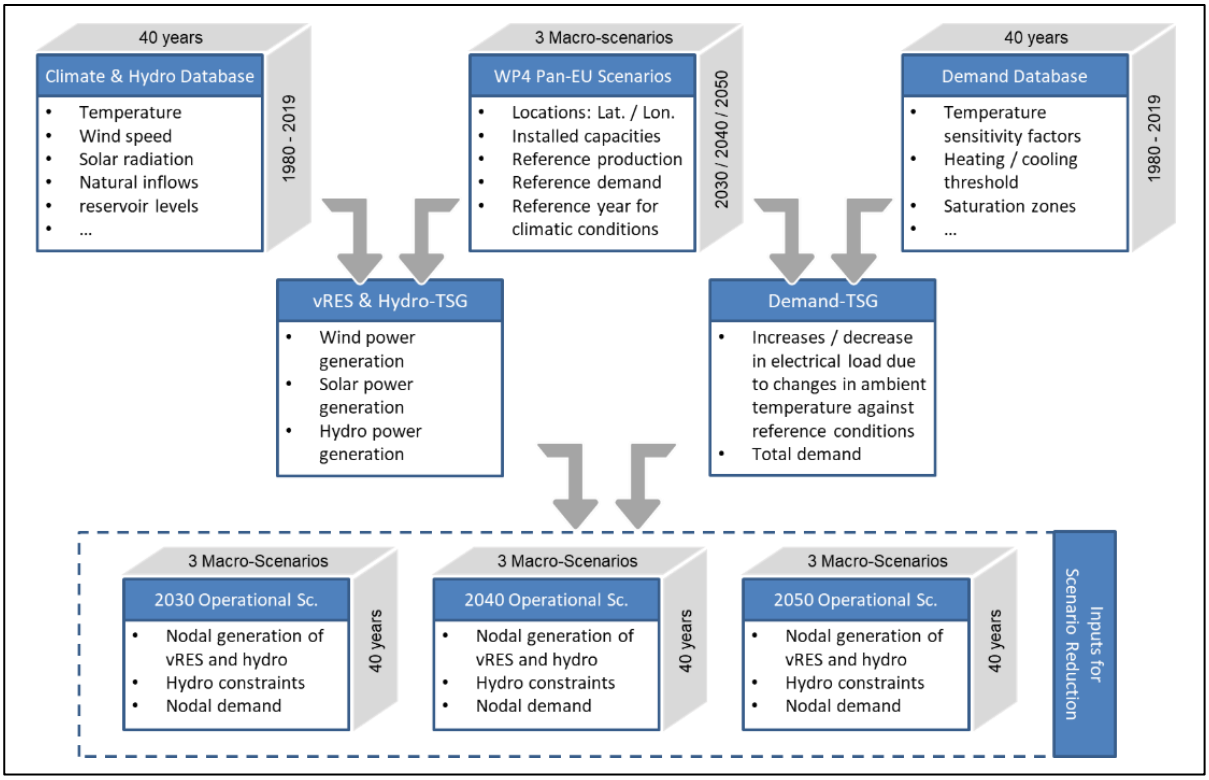

Figure 4-2: Flow chart of the developed methodology to generate operational scenarios

The generated sets of consecutive hourly time series data will finally be used as an input in the planning problem to determine the optimal grid expansion including storages and flexibility. As the amount of generated time series data (on nodal level) for uncertain input parameters becomes large, especially due to the consideration of various climatic conditions as well as the dynamic planning approach coupling the target years, a **scenario reduction** approach has to be applied to reduce the complexity. The scenario reduction methods are outlined in section 5. In the following sections, the developed scenario generation methodology is described.

4.1. Monte-Carlo Approach

The consideration of all stochastic inputs and their correlations respectively by an analytical approach is not compatible with a time series-based approach, i.e. the one adopted for the FlexPlan model. As such, ***a Monte Carlo method is applied to tackle the probabilistic interpretation*** (cf. section 3.1) ***of the underlying grid expansion planning problem*** including storage and flexibility.

A variety of samples, so called "**MC years**" can be used as an input for the underlying deterministic planning problem. Against the background of the deterministic grid expansion planning problem on an hourly basis for 8,760 consecutive time steps, the climatic and hydrological conditions as well as power plant availabilities for a full year are considered as stochastic inputs.

In this task, **three databases** have been set up for the sampling process. As mentioned before, these databases contain discrete time series data for the uncertain inputs namely vRES, hydro generation and load. There are no probability density functions used to replicate time series data or to generate samples.

The first database incorporates **climatic conditions** (in terms of hourly capacity factors) **for 40 years** (1980 – 2019) that are relevant to model vRES generation. The database also contains the ambient temperature which is used as a parameter to model variations of electrical load. A second database incorporates **hydrological conditions** (in terms of hourly capacity factors and normalized reservoir profiles) **for 36 years** (1982 – 2017) that are relevant to model the power output of run-of-river, reservoir and open-loop pumped storage plants. The last database contains parameters to model the s**ensitivity of the electrical load** as a function of the ambient temperature (with respect to a reference situation) for 1980 – 2019, as the first database covers exactly these 40 years. Additionally, the demand database includes normalized load profiles (in terms of hourly load factors) **for 35 years** (1982 – 2016). Based on the presented databases and the pool of time series data inside them, MC years can be sampled. Additionally, a fourth database containing **power plant outages** of WP 4 can be considered optionally.

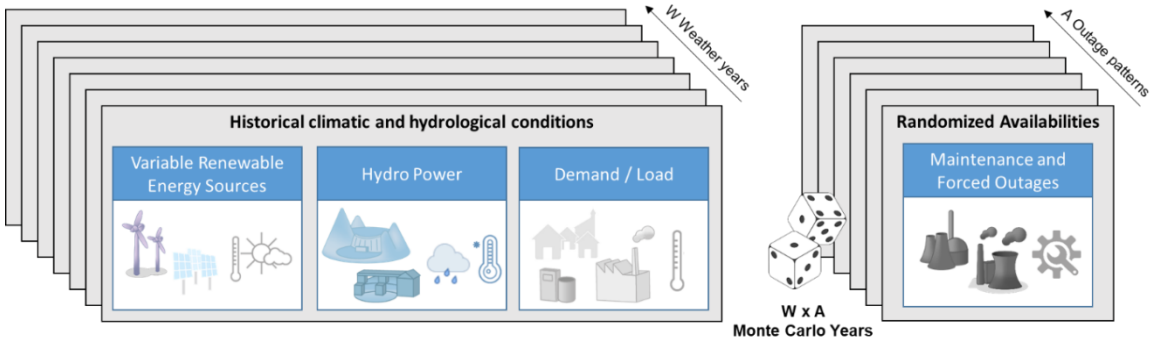

Figure 4-3: Monte Carlo scenario generation based on historical weather conditions and randomized availabilities

As the hydrological parameters published by ENTSO-E are generated by a statistical model taking into account different weather parameters of specific historical years, they are correlated with the climatic conditions in the respective historic year. ***Thus, climatic and hydrological conditions cannot be randomly combined to operational scenarios.*** Furthermore, uncertainties in the demand are correlated with wind, solar and hydro generation, due to their dependency on the temperature. ***Thus, only combinations of historical weather conditions (one set of correlated wind, solar and hydro as well as demand time series) and outages can be considered to generate various operational scenarios in terms of MC years.*** This process of randomly drawing one possible realization (historic year) from the pool of all realizations (total amount of historic observations) and subsequently combining the drawn realizations for vRES, hydro and load with a set of statistically independent power plant availabilities is called sampling. ***Thus, a MC year is one sample formed by a set of nodal time series describing one possible realization of the stochastic inputs on an hourly basis.***

The data sources as well as the specific methods applied to generate time series data while keeping spatial and temporal correlations is described in the following sections.

4.2. Modelling uncertainties in stochastic inputs

In this section, the considered stochastic inputs and the modelling of their respective uncertainties are presented. In a first step, the influencing factors of uncertainties are briefly outlined for each parameter considered within FlexPlan. Subsequently, the developed method to tackle the identified uncertainty as well as the approach for time series generation is presented. Finally, the methods used to incorporate spatial and temporal correlations are discussed.

4.2.1. Variable renewable energy sources

Renewable generation from vRES wind and solar power plants are treated as stochastic input parameters, as both generation technologies are heavily dependent on climatic conditions. With regard to wind power, the wind speed at hub-height at the plants location is the most influencing factor. For power generation of photovoltaic power plants, the solar irradiance at the plant's location as well as the orientation of a plant, i.e. of its solar PV panels, impacts its production significantly. Taking this into account, a database containing relevant climatic conditions to model vRES power output has been set up.

The **climatic database** contains bias-corrected hourly capacity factors for **wind and solar** per **NUTS-2-Region** for most of the European countries for (**1980 - 2019**) 40 historic years. In Figure 4-4 an overview of the modeled countries and their respective level of detail is provided.

Country	Country Code	Number of NUTS-2-Regions	Solar CF		Wind CF	
			NUTS-2-Level	Country	NUTS-2-Level	Country
AUSTRIA	AT	9	1980 - 2019	-	1980 - 2019	-
BELGIUM	BE	11	1980 - 2019	-	1980 - 2019	-
BOSNIA-HERZEGOVINA	BA	-	-	1980 - 2019		-
BULGARIA	BG	6	1980 - 2019	-	1980 - 2019	-
CROATIA	HR	2	1980 - 2019	-	1980 - 2019	-
CZECH REPUBLIC	CZ	8	1980 - 2019	-	1980 - 2019	-
DENMARK EAST	DK	5	1980 - 2019	-	1980 - 2019	-
DENMARK WEST	DK	5	1980 - 2019	-	1980 - 2019	-
ESTONIA	EE	1	1980 - 2019	-	1980 - 2019	-
FINLAND	FI	5	1980 - 2019	-	1980 - 2019	-
FRANCE	FR	27 (22)	1980 - 2019	-	1980 - 2019	-
GERMANY	DE	38	1980 - 2019	-	1980 - 2019	-
GREECE	EL	13	1980 - 2019	-	1980 - 2019	-
HUNGARY	HU	7	1980 - 2019	-	1980 - 2019	-
IRELAND	IE	2	1980 - 2019	-	1980 - 2019	-
ITALY	IT	21	1980 - 2019	-	1980 - 2019	-
LATVIA	LV	1	1980 - 2019	-	-	1980 - 2019
LITHUANIA	LT	1	1980 - 2019	-	-	1980 - 2019
LUXEMBOURG	LU	1	1980 - 2019	-	-	1980 - 2019
MACEDONIA	MK	1	1980 - 2019	-	-	1980 - 2019
MONTENEGRO	ME	1	-	1980 - 2019	-	-
NETHERLANDS	NL	12	1980 - 2019	-	1980 - 2019	-
NORWAY	NO	7	1980 - 2019	-	1980 - 2019	-
POLAND	PL	16	1980 - 2019	-	1980 - 2019	-
PORTUGAL	PT	7 (5)	1980 - 2019	-	1980 - 2019	-
ROMANIA	RO	8	1980 - 2019	-	1980 - 2019	-
SERBIA	RS	-	-	1980 - 2019	-	-
SLOVAKIA	SK	4	1980 - 2019	-	1980 - 2019	-
SLOVENIA	SI	2	1980 - 2019	-	1980 - 2019	-
SPAIN	ES	19 (18)	1980 - 2019	-	1980 - 2019	-
SWEDEN	SE	8	1980 - 2019	-	1980 - 2019	-
SWITZERLAND	CH	7	1980 - 2019	-	1980 - 2019	-
UNITED KINGDOM	UK	40	1980 - 2019	-	1980 - 2019	-

Figure 4-4: Overview of the climatic database for wind and solar capacity factors

The **capacity factor** is the ratio of realized generation to installed capacity in a specific period of time:

$$CF_{\Delta t}^{\text{Tech}} = \frac{E_{\Delta t}^{\text{Gen}}}{P^{\text{Inst}} \cdot \Delta t}$$

Thus, hourly capacity factors represent normalized generation profiles (in the interval [0, 1] per unit).

In Figure 4-5 the temporal and spatial diversity in wind generation over the past 40 years is depicted for two selected NUTS-2-regions in Germany. For the sake of visibility, the weekly moving average of the hourly time series data is plotted. Additionally, the 40 time series per region are plotted as fan charts to visualize the level of variability in historic data by providing the all-time median and percentiles respectively. The NUTS-2-region DEF0 is located in the north of Germany near the coast and DE21 is located in Bavaria in southern Germany (see Figure 2-3). As can be seen wind capacity factors are on average higher in the north and have a higher dispersion.

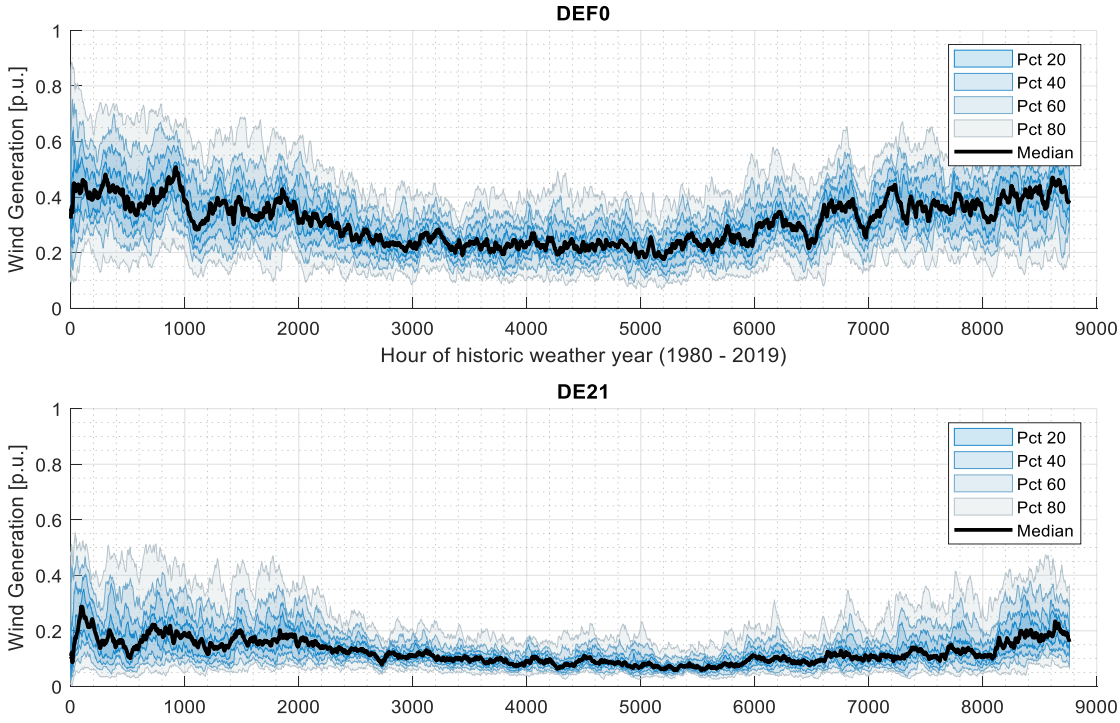

Figure 4-5: Exemplary visualization of wind capacity factors for two NUTS-2-regions in Germany (raw data taken from [96])

The key advantage of using capacity factors is that they can easily be used to calculate power generation time series considering different installed generation capacities, e.g. the ones coming from the pan -European scenarios.

The capacity factors used within FlexPlan have been calculated by the application of physical models for wind [47] and for PV [38] transforming numerical weather data in electrical outputs by applying the method of Staffel et al. [100] and Pfenninger et al. [38]. The data is publicly available at www.renewables.ninja [96]. In the following sections the models and processes applied by Staffel and Pfenninger to transform historical weather data of the MERRA-2 model in electrical power outputs of wind and solar power plants are presented.

Generating time series data for wind power generation

To generate wind energy time series the Virtual Wind Farm (VWF) Model [47] was applied by Staffel and Pfenninger. The basic functional principle of the applied physical model is shown in Figure 4-6.

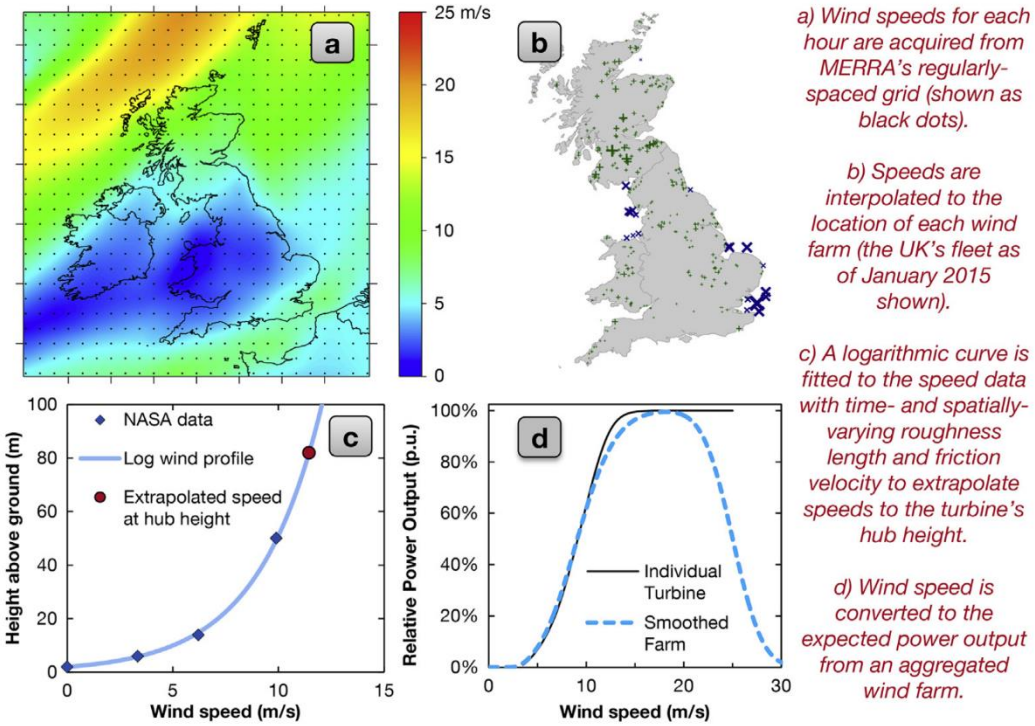

Figure 4-6: Overview of the VWF methodology (taken from [100])

The VWF model uses reanalysis data from the MERRA and MERRA-2 weather models. The freely available data of the models has a very good spatial and temporal resolution and is also available for a large time period. As shown in Figure 4-6, the VWF model first obtains historical wind speed data at different heights (2, 10 and 50 meters) above the ground for each grid point of the weather model (a). In a second step (b) the wind speeds of the grid points are interpolated to wind turbine sites using a local regression. Subsequently, the wind speed is extrapolated to hub height of the wind turbines for each site using the logarithmic profile law (c). Finally, the site-specific wind speeds are converted into power output using manufacturers' power curves (d). Smoothed power curves from manufacturers are used to account for the geographical mixing of a large number of different turbine types [100].

A complete mathematical description of the VWF model as well as the bias correction can be found in the supplementary material of [100].

Considering the simulated production of the wind farms $E_{\Delta t}^{\text{Wind,Sim}}$ and the implied installed wind capacity $P^{\text{Inst,Wind}}$, Staffel and Pfenninger derived hourly capacity factors $CF_{\Delta t}^{\text{Wind}}$ assuming that Δt is one hour:

$$CF_{\Delta t}^{\text{Wind}} = \frac{E_{\Delta t}^{\text{Wind,Sim}}}{P^{\text{Inst,Wind}} \cdot \Delta t} \quad \forall \Delta t \in TS_{\Delta t}$$

Finally, they assigned individual wind farms' data to NUTS-2-regions and calculated average wind capacity factors per NUTS-2-region for a wide range of historical weather conditions.

Staffel and Pfenninger compared their results with monthly average capacity factors of ENTSO-E to validate their results on country level. In Figure 4-7 one exemplary result (Germany 2005-2015) of their comparison using a calibrated VWF simulation model is depicted.

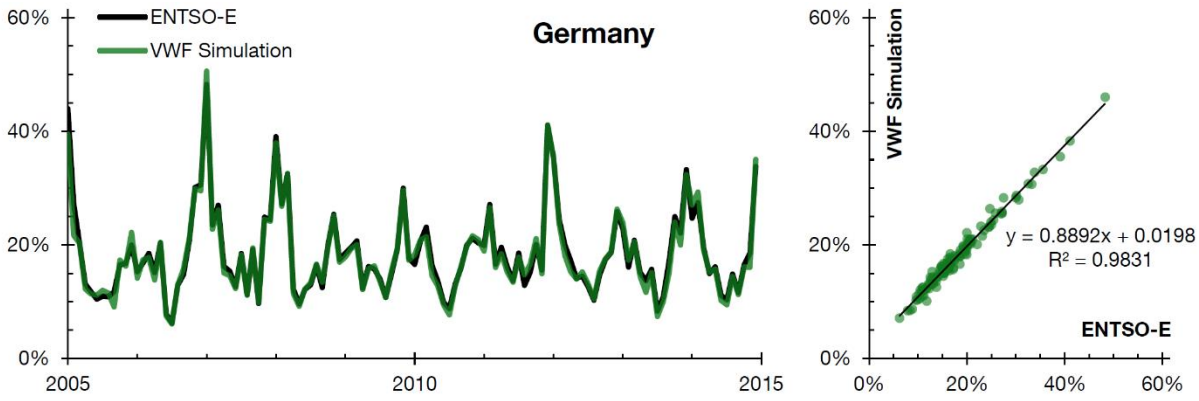

Figure 4-7: Monthly average capacity factors for Germany, comparing the calibrated VWF simulation (green) with ENTSO-E data (black) for 10 historic years (taken from supplementary material of [100])

As shown by Staffel and Pfenninger [100], the VWF simulation model is able to reproduce the intrinsic characteristics of wind power generation in Germany (and other European countries as well) on a monthly basis in the analyzed period of time. The correlation between the simulated and recorded values from ENTSO-E amounts to 98%. The authors of [100] have carried out extensive work to validate their results on an annual, monthly and hourly basis for most of the European countries. Furthermore, they have undertaken the seasonal and diurnal trends a detailed analysis om a country by country base which can be found in the supplementary material of [100]. In conclusion, it can be stated that performance of the VWF model is very well across Northwest Europe, with correlations to hourly capacity factors of above 0.90. However, Staffel and Pfenninger remark that the model's performance is noticeably worse in the Mediterranean, but still applicable for EU-wide analysis.

Thus, the hourly capacity factors derived by Staffel and Pfenninger (publicly available at www.renewables.ninja [96]) for 40 historical weather years on NUTS-2-Level are considered as adequate to model wind power generation in this project.

As can be seen in Figure 4-8, the installed wind capacities per node providing input to the scenario generation methodology are assigned to exactly one NUTS-2-region. The elementwise multiplication of the corresponding time series of hourly capacity factors for this NUTS-2-region with the nodal wind capacity of each node inside this NUTS-2-region provides the wind feed-in time series.

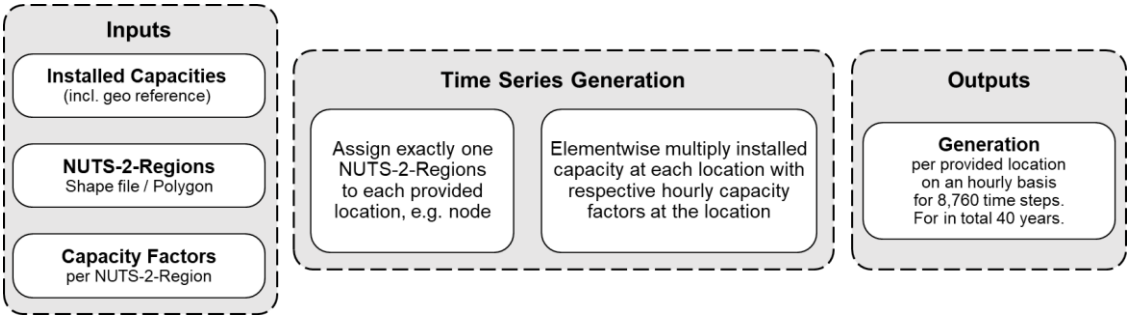

Figure 4-8: High level flow chart of the methodology for wind time series generation

Generating time series data for solar power generation

To transform solar irradiance data in PV power generation, the physical model Global Solar Energy Estimator (GSEE) [38] was applied by Pfenninger and Staffel. The basic working principle and data processing steps are presented in Figure 4-9.

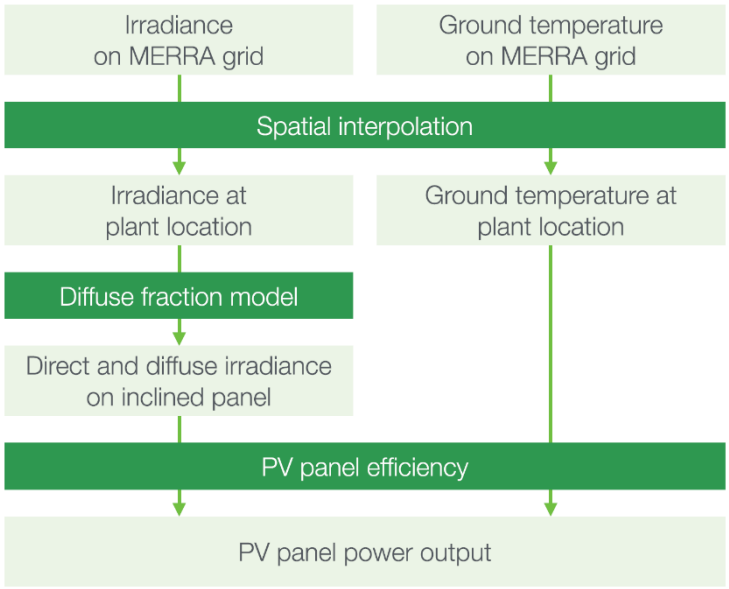

Figure 4-9: Overview of the approach used to model PV power output (taken from [38])

At first Pfenninger and Staffel carried out a spatial aggregation, as individual plant locations have to be assigned to a grid cell of the underlying MERRA-2 weather model. Based on the assigned weather data at the plant's location they approximated the diffuse irradiance fraction with the Boland-Ridley-Lauret (BRL) model [97] [98]. By the application of the BRL model, the direct and diffuse irradiances were calculated based on the weather data. Subsequently the irradiance on the plane of the PV panel was computed based on the direct and estimated diffuse irradiance. Thereby, tracking systems were not considered. Thus, Pfenninger and Staffel considered only fixed panels with a fixed azimuth angle as well as a fixed tilt angle. In a final step, the panel efficiency was taken into consideration to transform the solar irradiance into power outputs. Therefore, Pfenninger and Staffel applied a PV performance model [99], which provides temperature-dependent panel efficiency curves. Additionally, the losses inside the PV system, especially in the inverter converting the panel's DC output in AC power, are taken into consideration by a constant efficiency of 0.90 across all sites. A complete mathematical description of the GSEE model as well as the bias correction can be found in [38].

Considering the simulated PV production $E_{\Delta t}^{PV,Sim}$ and the implied installed PV capacity $P^{Inst,PV}$ Pfenninger and Staffel derived hourly capacity factors $CF_{\Delta t}^{PV}$ assuming that Δt is one hour:

$$CF_{\Delta t}^{PV} = \frac{E_{\Delta t}^{PV,Sim}}{P^{Inst,PV} \cdot \Delta t} \quad \forall \Delta t \in TS_{\Delta t}$$

Finally, they assigned individual grid cells to NUTS-2-regions and calculated average PV capacity factors per NUTS-2-region for a wide range of historical weather conditions.

In Figure 4-10 the temporal and spatial diversity in solar generation over the past 40 years is depicted for two NUTS-2-regions in Germany. For sake of visibility, the weekly moving average of the hourly time series data is plotted. Additionally, the 40 time series per region are plotted as fan charts to visualize the level of uncertainty in historic data by providing the all-time median and percentiles respectively. The NUTS-2-region DE50 is located in the north of Germany near the coast and DE27 is located in Bavaria in southern Germany (see Figure 2-3). As can be seen, solar capacity factors across Germany are quite similar. In comparison to wind capacity factors the solar generation is heavily dependent on the season.

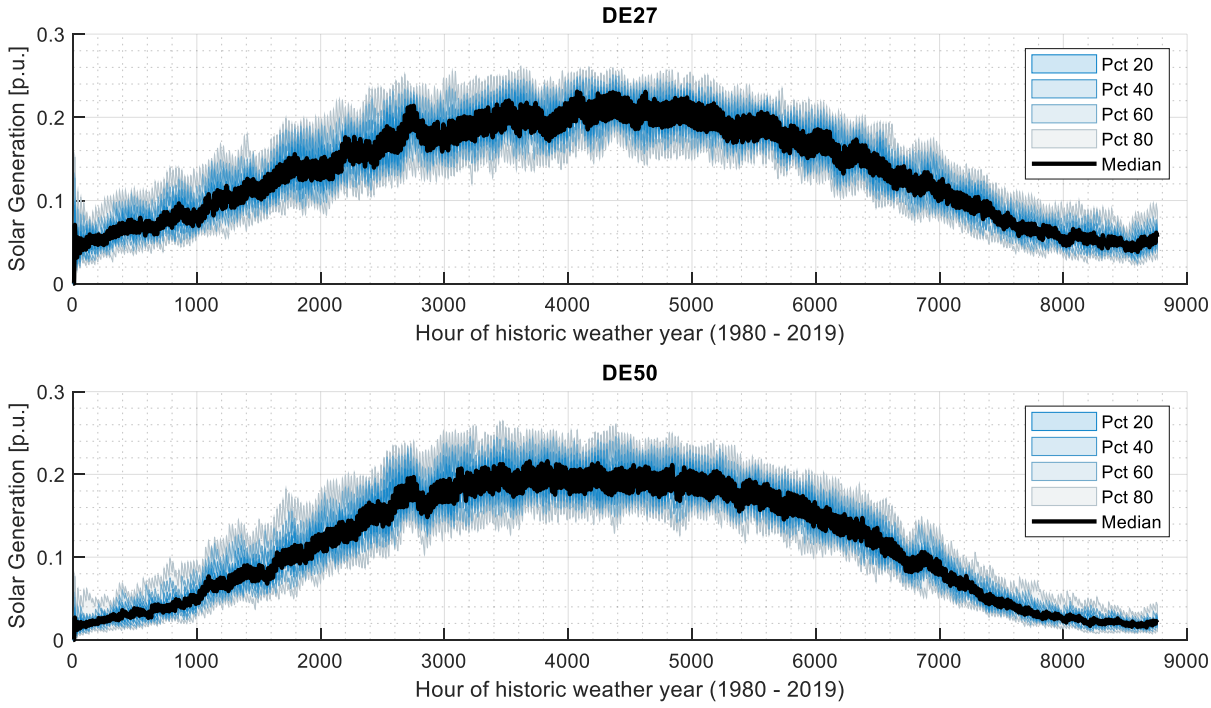

Figure 4-10: Exemplary visualization of solar capacity factors for two NUTS-2-regions in Germany (raw data taken from [96])

Pfenninger and Staffel compared their results with weekly average capacity factors of ENTSO-E to validate their results on country level. In Figure 4-11 one exemplary result (Germany 2014) of their comparison using their calibrated GSEE simulation model is depicted.

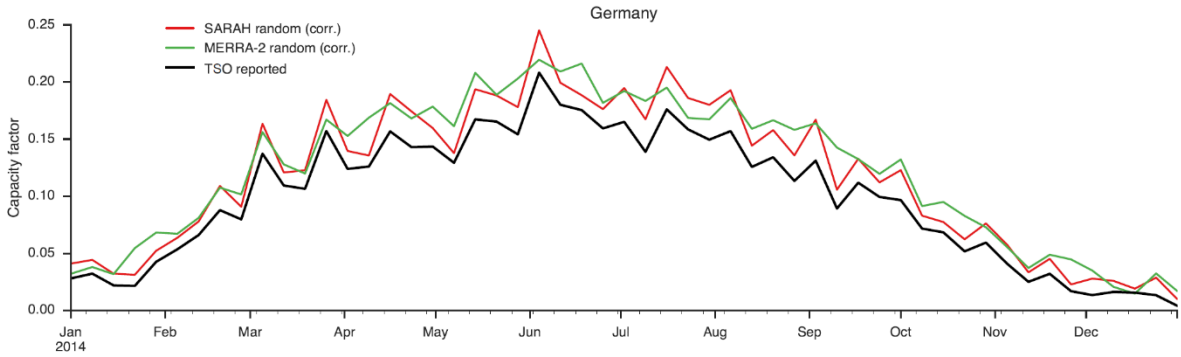

Figure 4-11: Overview of the approach used to model PV power output (taken from [38])

As shown by Pfenninger and Staffel [38] the GSEE simulation model is able to reproduce the intrinsic characteristics of solar power generation in Germany (and other European countries as well) on a weekly basis in the analyzed year. The authors of [38] have carried out extensive work to validate their results on an annual, monthly and hourly basis for most of the European countries. Furthermore, they have undertaken the seasonal and diurnal trends a detailed analysis country by country that can be found in the supplementary material of [38]. In conclusion, it can be stated that performance of the GSEE model is suitable for the approach taken by FlexPlan. However, Pfenninger and Staffel remark that the model tends to overestimate PV generation in the southern EU countries, but still being applicable for EU-wide analysis.

Thus, **it can be concluded that** the hourly capacity factors derived by Pfenninger and Staffel (publicly available at www.renewables.ninja [96]) for 40 historical weather years on NUTS-2-Level are considered as adequate to model solar power generation in this project.

Similar to the generation of wind power time series (cf. Figure 4-8), the installed PV capacities per node provided are assigned are assigned to exactly one NUTS-2-region. As a consequence, the corresponding time series of hourly generation is obtained by elementwise multiplication of the capacity factors for this NUTS-2-region with the nodal PV capacity of each node inside this NUTS-2.

Modelling spatial correlations

The climatic database created in this task contains **capacity factors per NUTS-2-Region**, i.e. characteristic differences in wind and solar generation due to heterogeneous weather conditions are considered in detail. The following Figure 4-12compares NUTS-2-Regions and transmission grid nodes modelled in Germany.

As can be seen from Figure 4-12 the number of transmission grid nodes exceeds the number of NUTS-regions significantly. It should be noted that this is the case for nearly all countries modelled within FlexPlan. Bosnia Herzegovina, Montenegro, North Macedonia and Luxembourg are exception as they do not have more than one NUTS-2-region. As a consequence, an intermediate step has to be carried out to generate nodal time series data based on NUTS-2-level capacity factors. The intermediate step includes the **allocation of nodal installed vRES capacity** information **to NUTS-2-Regions** considered in in this task. Thus, for each modelled node and NUTS-region it is checked whether a node is located inside a NUTS-2 region. As a result, each modelled transmission grid node $n \in S_n$ is assigned to exactly one NUTS-region $r \in S_r$ such that the installed wind and solar capacity per NUTS-2-region $P_{r,m}^{\text{Inst}}$ can be calculated as the sum of its assigned grid nodes for one specific macro-scenario $m \in S_m$:

$$P_{r,m}^{\text{Inst}} = \sum_{n \in r} P_{n,m}^{\text{Inst}} \quad \forall r \in RS_r$$

After calculating the installed capacity per NUTS-region for one specific macro-scenario $P_{r,m}^{\text{Inst}}$, the nodal time series of power generation for wind and solar for one operational scenario $s \in S_s$ is determined by multiplying the installed capacity with the technology specific hourly capacity factors $CF_{r,t,s}^{\text{Tech}}$ (for wind and solar) in the respective NUTS-region r and in each time step t:

$$P_{r,t,m,s}^{\text{Gen}} = P_{r,m}^{\text{Inst}} \cdot CF_{r,t,s}^{\text{Tech}} \quad \forall r \in R, S_r, \forall t \in TS_t$$

To put in a nutshell, individual climatic conditions are considered per NUTS-2-Region to model spatial correlations in wind and solar power generation all over Europe. For sake of flexibility and applicability of the scenario and time series generation method, climatic conditions are indirectly modeled in terms of bias-

corrected hourly capacity factors that are easily merged with installed capacities provided as input. However, it has to be noted that this approach will lead to identical generation profiles with respect to the shape at all nodes that are assigned to the same NUTS-2-Region. This might be a problem for very detailed studies, but should not be a major drawback in Pan-EU expansion planning, as large-scale spatial correlations are considered and NUTS-2-Level can be considered as adequate for transmission grid studies.

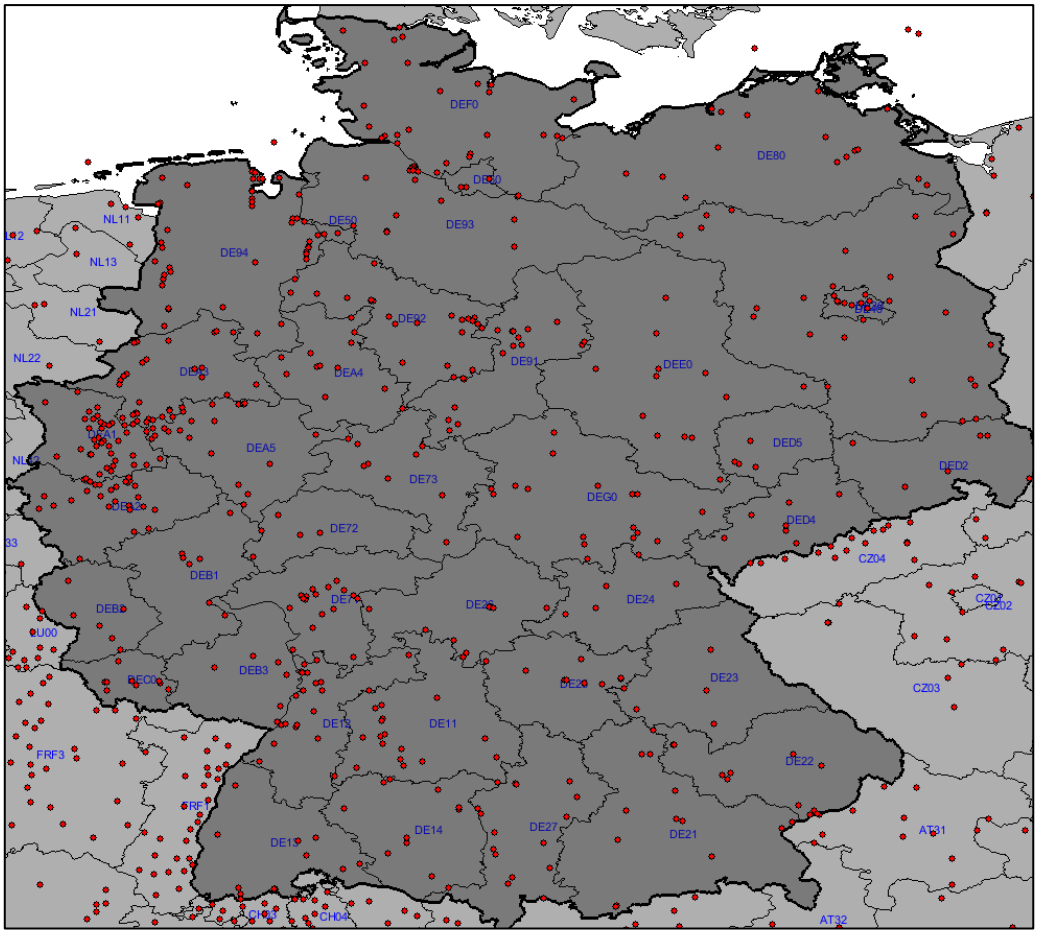

Figure 4-12: Comparison of NUTS-2-Regions with modelled transmission grid nodes in Germany

4.2.2. Hydro power

Run of River (RoR) and **reservoir generation** as well as **natural inflows** are treated as stochastic inputs to the planning problem, as these parameters are heavily dependent on hydrological conditions. In contrast to wind and solar power generation it is important to note that hydro generation cannot be directly derived from climatic conditions with physical models. On the one hand, the power output hydro power plants depends on the availability of water, i.e. natural inflows, just as wind and solar power plants are dependent on climatic conditions. However, hydro power plants are able to store water in reservoirs enabling them to be dispatchable within certain limits, which clearly distinguishes them from non-dispatchable vRES. On the other hand, hydro power plants have to consider additional operational limits, e.g. minimum and maximum reservoir levels and generation, restricting their operational flexibility.

With regard to RoR generation the minimum throughput, i.e. the flow rate, is the most influencing factor causing a non-flexible minimum production capacity leading to a partly non-dispatchable power output.

Reservoir plants' as well as open-loop pumped storage plants' generation is impacted by natural inflows as well as their respective (initial) storage capacities significantly. Taking this into account, a database containing relevant hydrological conditions and parameters to model hydro generation has been set up.

The **hydrological database** contains hourly capacity factors for RoR generation, normalized minimum and maximum RoR and reservoir generation, normalized minimum and maximum reservoir levels and cumulated natural inflows for reservoir and open-loop pumped storage plants for most of the European countries for (**1982 - 2017**) 36 historic years. The hydrological database's spatial level of detail is depicted in Figure 4-13.

Country	Country Code	Market Zones	Hydro RoR CF
ALBANIA	AL		1982 - 2017
AUSTRIA	AT		1982 - 2017
BOSNIA-HERZEGOVINA	BA		1982 - 2017
BELGIUM	BE		1982 - 2017
BULGARIA	BG		1982 - 2017
SWITZERLAND	CH		1982 - 2017
CZECH REPUBLIC	CZ		1982 - 2017
GERMANY	DE		1982 - 2017
SPAIN	ES		1982 - 2017
FRANCE	FR		1982 - 2017
GREECE	GR		1982 - 2017
CROATIA	HR		1982 - 2017
HUNGARY	HU		1982 - 2017
IRELAND	IE		1982 - 2017
ITALY	IT	6	1982 - 2017
LITHUANIA	LT		1982 - 2017
LUXEMBOURG	LU	2	1982 - 2017
LATVIA	LV		1982 - 2017
MONTENEGRO	ME		1982 - 2017
MACEDONIA	MK		1982 - 2017
NETHERLANDS	NL		1982 - 2017
NORWAY	NO	3	1982 - 2017
POLAND	PL		1982 - 2017
PORTUGAL	PT		1982 - 2017
ROMANIA	RO		1982 - 2017
SERBIA	RS		1982 - 2017
SWEDEN	SE	4	1982 - 2017
SLOVENIA	SI		1982 - 2017
SLOVAKIA	SK		1982 - 2017
TURKEY	TR		1982 - 2017
UNITED KINGDOM	UK		1982 - 2017

Figure 4-13: Overview of the hydrological database for RoR and reservoir capacity factors

The initial daily and weekly absolute values of realized generation as well as minimum and maximum generation have been calculated by Milano Multiphysics S.r.l.s. for the ENTSO-E using a statistical model. A transfer function was built by correlating historical water volumes flowing in rivers with the corresponding hydro power generation for a number of sample years. ENTSO-E published the dataset including a short documentation [101] as part of its Summer Outlook 2020 [102].

Generating time series data

Similar to wind and PV capacity factors, hourly RoR capacity factors representing normalized generation profiles (in the interval [0, 1]) have been calculated based on ENTSO-E's initial dataset including absolute values for hydro power. Hydro capacity factors are calculated by dividing the realized generation $E_{\Delta t}^{\mathrm{Gen}}$ (taken from ENTSO-E's new hydro dataset) by the installed reference capacity P^{Inst} (provided in ENTSO-E's new hydro dataset as well) multiplied with the specific observation period Δt taking into account that part of the published data is either some parameters on a daily or weekly basis:

$$CF_{\Delta t}^{\mathrm{Tech}} = \frac{E_{\Delta t}^{\mathrm{Gen}}}{P^{\mathrm{Inst}} \cdot \Delta t}$$

For sake of consistency, within this task, daily and weekly time series data were interpolated to hourly values that are in line with the temporal resolution of the FlexPlan planning tool.

The statistical method that was developed and applied by Milano Multiphysics for ENTSO-E to generate time series of the electrical output of hydro power plants and their respective operational constraints is based on machine learning (neural networks) as there is no way of applying physical models in the field of hydro power yet. The statistical model makes use of the total unregulated inflows that were calculated by the European Hydrological Predictions for the Environment and consists of a high resolution (E-HYPE) [103] model of the Swedish Meteorological and Hydrological Institute (SMHI). Furthermore, a number of additional climatic parameters based on reanalysis from Clim4Energy – Copernicus Project [104] were used by Milano Multiphysics as an input for the statistical model. Figure 4-14 visualizes the correlation between inflows per catchment hydro power plant's production.

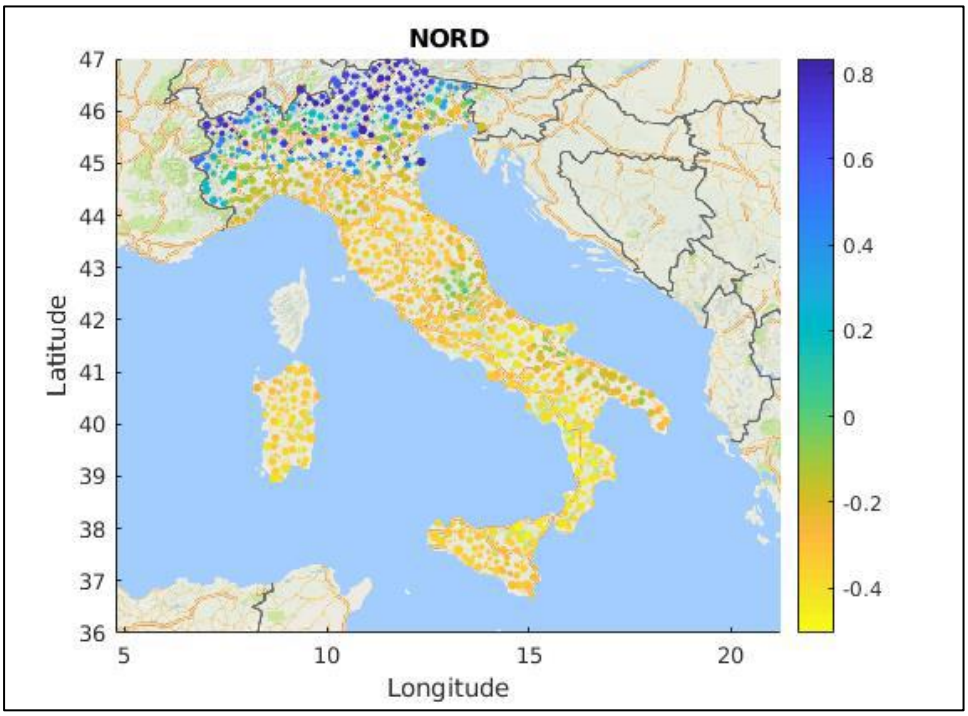

Figure 4-14: Correlation between inflows per catchment and hydro power plant's production (taken from [101])

Besides the hydrological and climatic data, detailed technical information and historical production data of hydro power plants for eight historic years have been collected on a unit-by-unit basis by ENTSO-E and provided to Milano Multiphysics as training data for the statistical machine learning model.

The basic idea is to use historic observations (2010 - 2017) of hydro power generation in GWh per day from ENTSO-E and SMHI E-HYPE reanalysis data (1982 – 2017) of unregulated inflows in m^3 per day to derive a transfer function (m^3/d → GWh/d) that can be used to calculate hydro generation based on historical inflows for periods outside the training period. The working principle of the statistical model applied by Milano Multiphysics can be summarized as follows:

1. Time series of SMHI's total unregulated inflows were normalized to obtain zero-mean and unit standard deviation distributions to ensure an optimal data decomposition in the following dimensionality reduction phase.

2. Singular-value decomposition for dimensionality reduction by orthogonal decomposition of SMHI inflow data to avoid an overfitting of the regression and improve computational performance.
3. Estimation of transfer functions through regression by least squares plant by plant for SMHI's inflows data and respective hydro power plant's observed generation.
4. Post-processing of production data in its spectrum by Fourier transformation analyzing intertemporal dynamics of pure and regulated RoR plants.
5. Zonal aggregation and validation of results with an independent historic input data set that has not been used in the training process to check if the model tends to overfit.

In Figure 4-15 Milano Multiphysics' modeled and real RoR generation in Italy (North) between 2010 and 2016 is compared. Additionally, the resulting hydro generation from 1981 to 2016 of Milano Multiphysics' machine learning approach based on ENTSO-E's provided training data is visualized. The statistical model performs quite well and is able to generate hydro generation time series data for various historic years using the identified transfer functions. Thus, the ENTSO-E hydro data set is used as basis for hydro time series generation in this project.

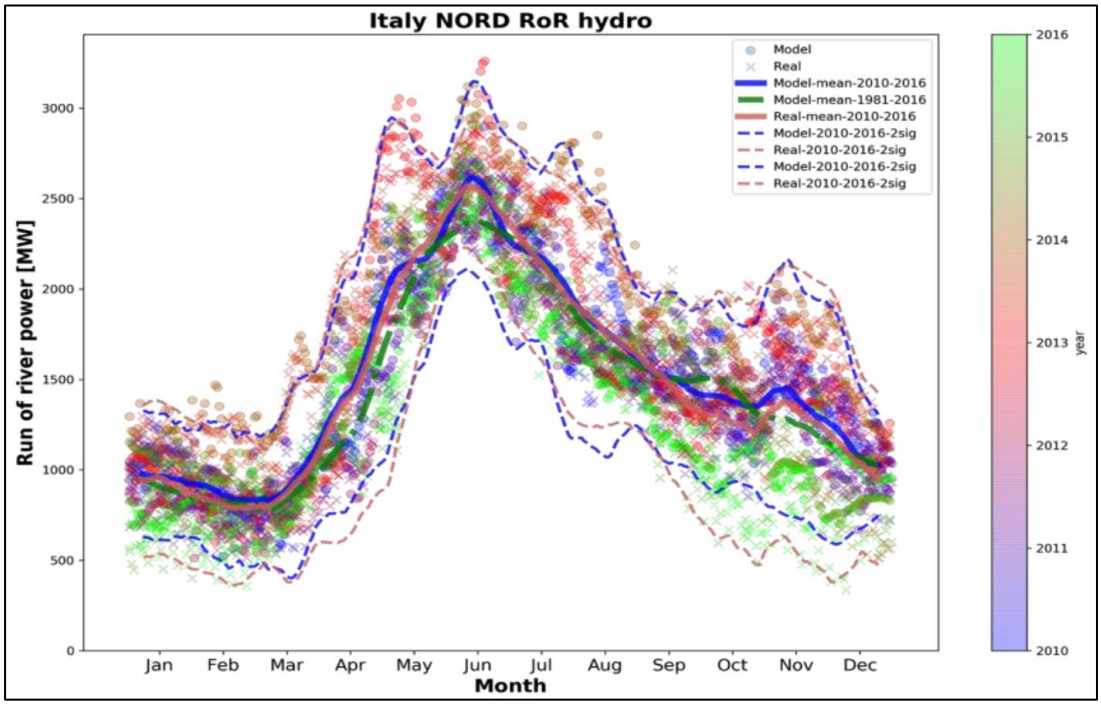

Figure 4-15: Comparison of modelled and real hydro generation on national level (taken from [101])

Summarizing, the ENTSO-E dataset provides the following **outputs per country for 36 years**:

- Run-of-River and Pondage:
  - Reference installed capacity (MW)
  - Reference storage capacity (MWh)
  - Daily absolute generation (GWh)
  - Daily min. / max. generation constraints (MW)
- Reservoirs:
  - Reference installed capacity (MW)

- Reference storage capacity (MWh)
- Weekly min. / max. reservoir levels (p.u.)
- Weekly min. / max. generation constraints (MW)
- Weekly cumulated inflows (GWh)

- Open Loop Pump Storages:
  - Reference installed capacity (MW)
  - Reference storage capacity (MWh)
  - Weekly min. / max. reservoir levels (p.u.)
  - Weekly min. / max. generation constraints (MW)
  - Weekly cumulated inflows (GWh)

In Figure 4-16 the distribution of the daily RoR generation is depicted exemplary for Italy (North).

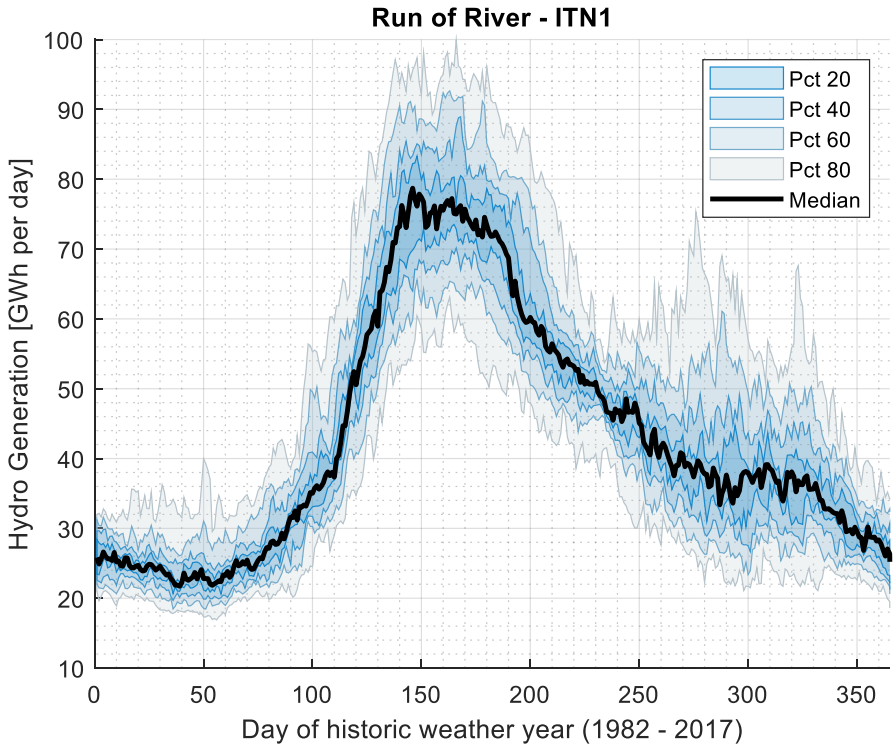

Figure 4-16: Distribution of daily run of river generation for 36 historical years in Italy North (raw data taken from ENTSO-E [102])

As can be seen from Figure 4-16 the median of the aggregated hydro generation by RoR plants has a distinctive seasonal pattern. It has to be noted that the pattern is different for each country due to different heterogeneous factors influencing hydro generation in general. Considering the visualized percentiles, the level of uncertainty in hydro generation over time becomes directly visible.

Within this task, firstly, the absolute values taken from the ENTSO-E dataset were normalized considering the provided reference capacities. Secondly, daily and weekly parameters were interpolated to obtain an hourly time series to be in line with the other inputs of the planning tool. Subsequently, the hourly hydro capacity factors $CF_{c,t,s}^{\mathrm{HY,RoR}}$ for a country $c \in S_c$ is used to calculate the nodal hourly hydro RoR generation in each time step $t \in S_t$ for one operational scenario $s \in S_s$ considering the nodal installed capacities of hydro RoR $P_{n,m}^{\mathrm{HY,RoR,Inst}}$ of one specific macro-scenario $m \in S_m$ as an input:

$$P_{n,t,m,s}^{\mathrm{HY,Gen}} = P_{n,m}^{\mathrm{HY,RoR,Inst}} \cdot CF_{c,t,s}^{\mathrm{HY,RoR}} \; \forall n \in N, S_n, \forall t \in TS_t$$

The methodology applied in this task for hydropower time series generation is depicted in Figure 4-17. The installed hydropower capacities per node (differentiated by hydro type) provided by *MILES* as input are assigned to exactly one ENTSO-E market area. As a consequence, the corresponding time series of hourly capacity factors for this market area is elementwise multiplied with the nodal hydro capacity of each node inside this area to generate hydro RoR feed-in time series.

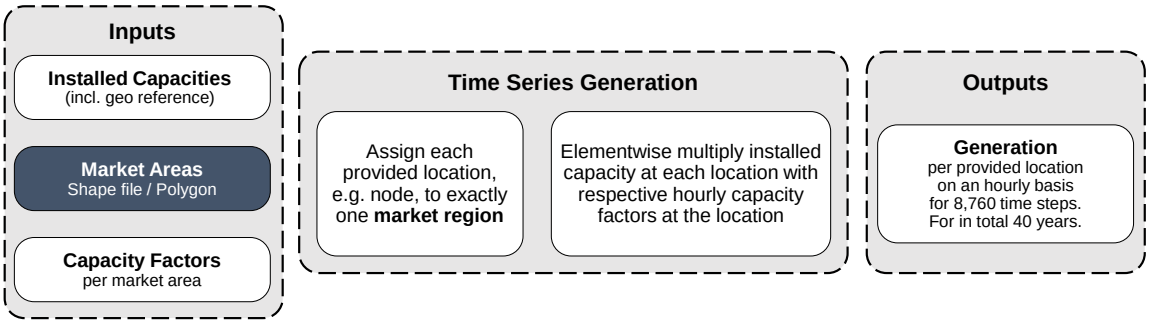

Figure 4-17: High level flow chart of methodology for hydro time series generation

Modelling spatial correlations

The hydrological database created in this task contains **capacity factors per country** and in some cases per ENTSO-E market zone if a country contains more than one market zone, i.e. characteristic differences in hydro generation due to heterogeneous climatic and hydrological conditions are considered less detailed in comparison to vRES generation on NUTS-2-Level. However, individual hydro conditions in terms of generation profiles and reservoir levels shaping total hydro generation are considered per country. Thus, the individual intrinsic characteristics of hydro generation in Scandinavia and the Alpine region are considered as well as possible considering the amount of available data as well as its level of detail.

4.2.3. Load

With regard to the demand side, the electrical load demand is treated as a stochastic input parameter in FlexPlan's new planning tool, as it is heavily dependent on climatic conditions. The outdoor ambient temperature is the most influencing factor, as electrical heating is widely used in the commercial and domestic sector. The additional volatile electrical demand for heating and cooling changes throughout the year (seasonality). Additionally, extreme events, such as colds spells in winter and heat waves / droughts in summer impact the peak load and thereby the load profiles during such kind of extreme events. With these aspects in mind, a database containing load-temperature sensitivity factors has been set up.

The **demand sensitivity database** depicted in Figure 4-18 contains in total six parameters to model the **load-temperature sensitivity** for most of the European countries by characteristic curves derived from regression analyses that have been carried out by the ENTSO-E.

Country	Country Code	Market Zones	Load CF
ALBANIA	AL		1982 - 2016
AUSTRIA	AT		1982 - 2016
BOSNIA-HERZEGOVINA	BA		1982 - 2016
BELGIUM	BE		1982 - 2016
BULGARIA	BG		1982 - 2016
SWITZERLAND	CH		1982 - 2016
CYPRUS	CY		1982 - 2016
CZECH REPUBLIC	CZ		1982 - 2016
GERMANY	DE		1982 - 2016
DENMARK EAST	DKE		1982 - 2016
DENMARK WEST	DKW		1982 - 2016
ESTONIA	EE		1982 - 2016
SPAIN	ES		1982 - 2016
FINLAND	FI		1982 - 2016
FRANCE	FR	2	1982 - 2016
GREECE	GR	2	1982 - 2016
CROATIA	HR		1982 - 2016
HUNGARY	HU		1982 - 2016
IRELAND	IE		1982 - 2016
ITALY	IT	6	1982 - 2016

Country	Country Code	Market Zones	Load CF
LITHUANIA	LT		1982 - 2016
LUXEMBOURG	LU	3	1982 - 2016
LATVIA	LV		1982 - 2016
MONTENEGRO	ME		1982 - 2016
MACEDONIA	MK		1982 - 2016
MALTA	MT		1982 - 2016
NETHERLANDS	NL		1982 - 2016
NORWAY	NO	3	1982 - 2016
POLAND	PL		1982 - 2016
PORTUGAL	PT		1982 - 2016
ROMANIA	RO		1982 - 2016
SERBIA	RS		1982 - 2016
SWEDEN	SE	4	1982 - 2016
SLOVENIA	SI		1982 - 2016
SLOVAKIA	SK		1982 - 2016
TURKEY	TR		1982 - 2016
UKRAINE	UA		1982 - 2016
UNITED KINGDOM	UK		1982 - 2016
NORTHERN IRELAND	NI		1982 - 2016

Figure 4-18: Overview of the demand database for load capacity factors and sensitivity factors

The parameters used to model the electrical loads' sensitivity with respect to the outdoor temperature are stylized in Figure 4-19. For comparably low temperatures beneath 15°C, the electrical energy consumed for heating increases significantly (Heating Zone). If outdoor temperature's daily average remains between +15°C and +20°C, neither heating nor cooling is used. Thus, no additional energy is consumed by heating and cooling devices (Comfort Zone). If the outdoor temperature rises beyond 21°C, cooling devices are turned on (Cooling Zone). As a consequence, the amount of additional energy used for cooling increases as well. Taking into consideration that heating and cooling zones have limits, due to saturation effects lower and upper bounds are defined respectively (Saturation Zones).

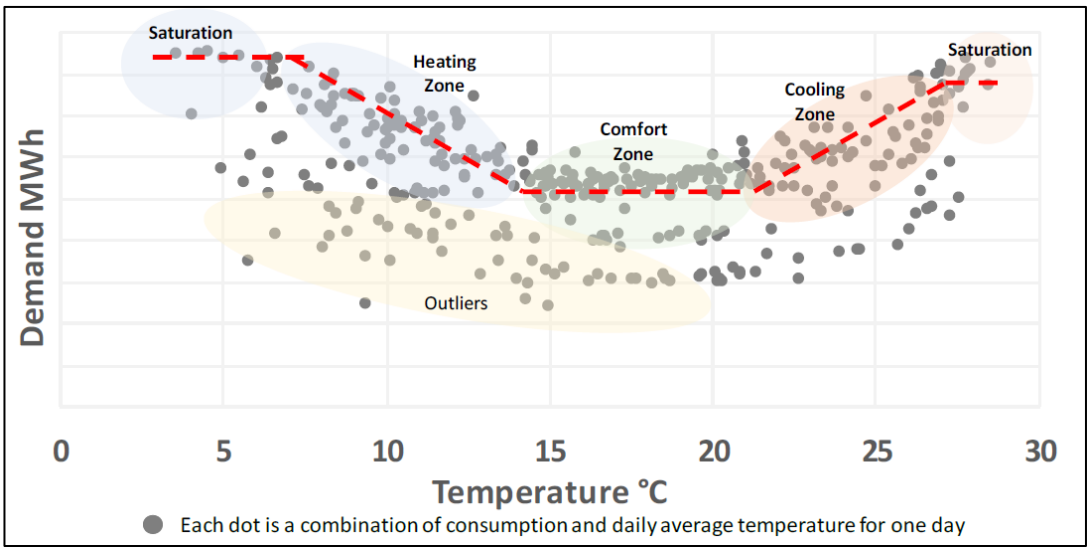

Figure 4-19: Schematic visualization of load-temperature sensitivity by: heating, comfort, cooling, and saturation zones that are used to form the characteristic curves and its parameters (taken from [105])

Due to individual heating concepts across Europe, e.g. France vs. Germany, it is obvious that the following parameters defining the presented zones differ from country to country significantly:

- [- Infinity, heating zone saturation temperature] (Saturation)
- [heating zone saturation temperature, heating zone limit temperature] (Heating zone)
- Load-temperature sensitivity factor inside heating zone LF_c^{Heat} [MW/°C]
- [heating zone limit temperature, cooling zone limit temperature] (Comfort zone)
- [cooling zone limit temperature, cooling zone saturation temperature] (Cooling zone)
- Load-temperature sensitivity factor inside cooling zone LF_c^{Cool} [MW/°C]
- [cooling zone saturation temperature, + Infinity] (Saturation)

To approximate the additional demand due to deviations in climatic conditions the so-called **load-temperature sensitivity factors** (LF_c^{Heat} and LF_c^{Cool}) are used, **defining the line's slope**, as can be seen from Figure 4-19. The key advantage of these characteristic curves is that they can easily be used to calculate load time series considering a reference demand time series (and information on the assumed reference year) is provided, e.g. the ones coming from *MILES*. The load-temperature sensitivity factors themselves have been calculated by ENTSO-E applying a regression model and published in [106].

Generating time series data

For the generation of hourly load time series data, information from different internal (climatic database and demand coefficients database) and external sources (initial nodal load time series as of WP 4) are merged. The initial load time series calculated with *MILES* are based on macro-assumptions concerning the future energy system as well as one specific historical reference year. Thus, it is essential to **match the reference year** used in *MILES* with the reference year used in this tasks' scenario and time series generation approach, as the change in load due to deviations in the ambient temperature is calculated as a time series of differences.

In Figure 4-20 the distribution of historical hourly temperature values (2m over ground) is depicted for Germany, as an exemplary excerpt of the developed climatic database.

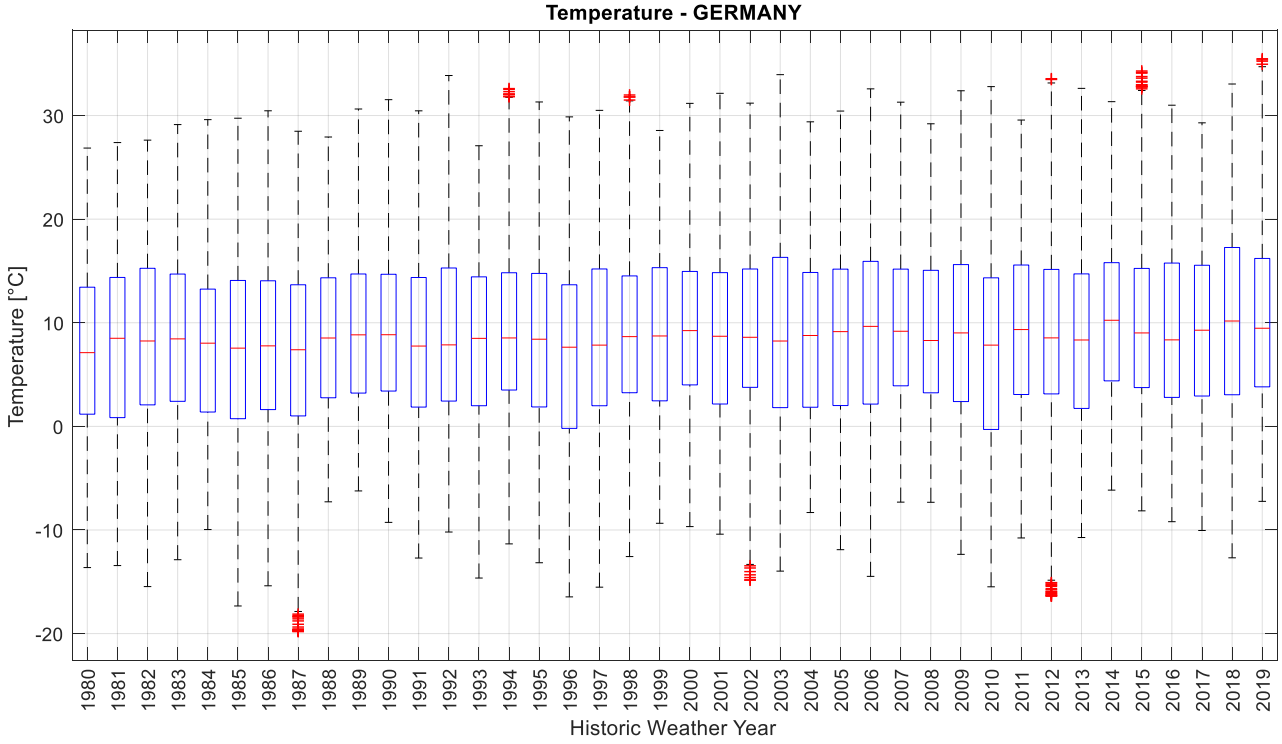

Figure 4-20: Boxplot of historical ambient temperatures in Germany from 1980 – 2019 (raw data taken from [96])

Considering that the hourly load time series of one specific macro-scenario that has been generated based on reference climatic conditions, e.g. historic observations from 2012, the climatic conditions $s_{\mathrm{Ref}} \in S$ from the climatic database are considered as a reference in the scenario generation approach. Subsequently, another historic observation $s \in S_s$ is randomly drawn from the climatic database and used to calculate the absolute hourly deviation ($\Delta\vartheta^{\mathrm{Pos}}_{c,t,s,s_{\mathrm{Ref}}}$ and $\Delta\vartheta^{\mathrm{Neg}}_{c,ts,s_{\mathrm{Ref}}}$) of the ambient temperature for a country $c \in S_c$ taking into account the country specific heating, cooling, comfort and saturation zones (cf. Figure 4-19). Once the directional differences are calculated for each hour of the year, the increase/decrease ($\Delta P^{\mathrm{Load,Heat}}_{c,t}$ and $\Delta P^{\mathrm{Load,Cool}}_{c,t}$) with respect to the reference year is calculated considering the zones' specific slopes (LF^{Heat}_{c} and LF^{Cool}_{c}):

$$\Delta P^{\mathrm{Load,Heat}}_{c,t,s,s_{\mathrm{Ref}}} = LF^{\mathrm{Heat}}_{c} \cdot \Delta\vartheta^{\mathrm{Pos}}_{c,t,s,s_{\mathrm{Ref}}} \; \forall t \in T$$

$$\Delta P^{\mathrm{Load,Cool}}_{c,t,s,s_{\mathrm{Ref}}} = LF^{\mathrm{Cool}}_{c} \cdot \Delta\vartheta^{\mathrm{Neg}}_{c,t,s,s_{\mathrm{Ref}}} \; \forall t \in T$$

In Figure 4-21 the temporal and spatial diversity (only due to temperature-sensitivity) in electricity demand over the past 40 years is depicted for Italy's six market zones. For the sake of visibility, only a limited number of 400 time steps (hours 5,400 to 5,800 of the year) is plotted. Additionally, the 40 time -series per region are plotted as fan charts to visualize the level of uncertainty in historic data as percentiles. There are six market areas in Italy. Basically, basically dividing the country in four zones: North (ITN1), Central North (ITCN), Central South (ITCS), South (ITS1) as well as two islands Sardinia (ITSA) and Sicily (ITSI) forming two additional zones. As can be seen from Figure 4-21, the hourly load capacity factors in the different regions have their individual characteristics. Furthermore, the level of uncertainty due to temperature-sensitive loads is less prominent in the ITSA zone.

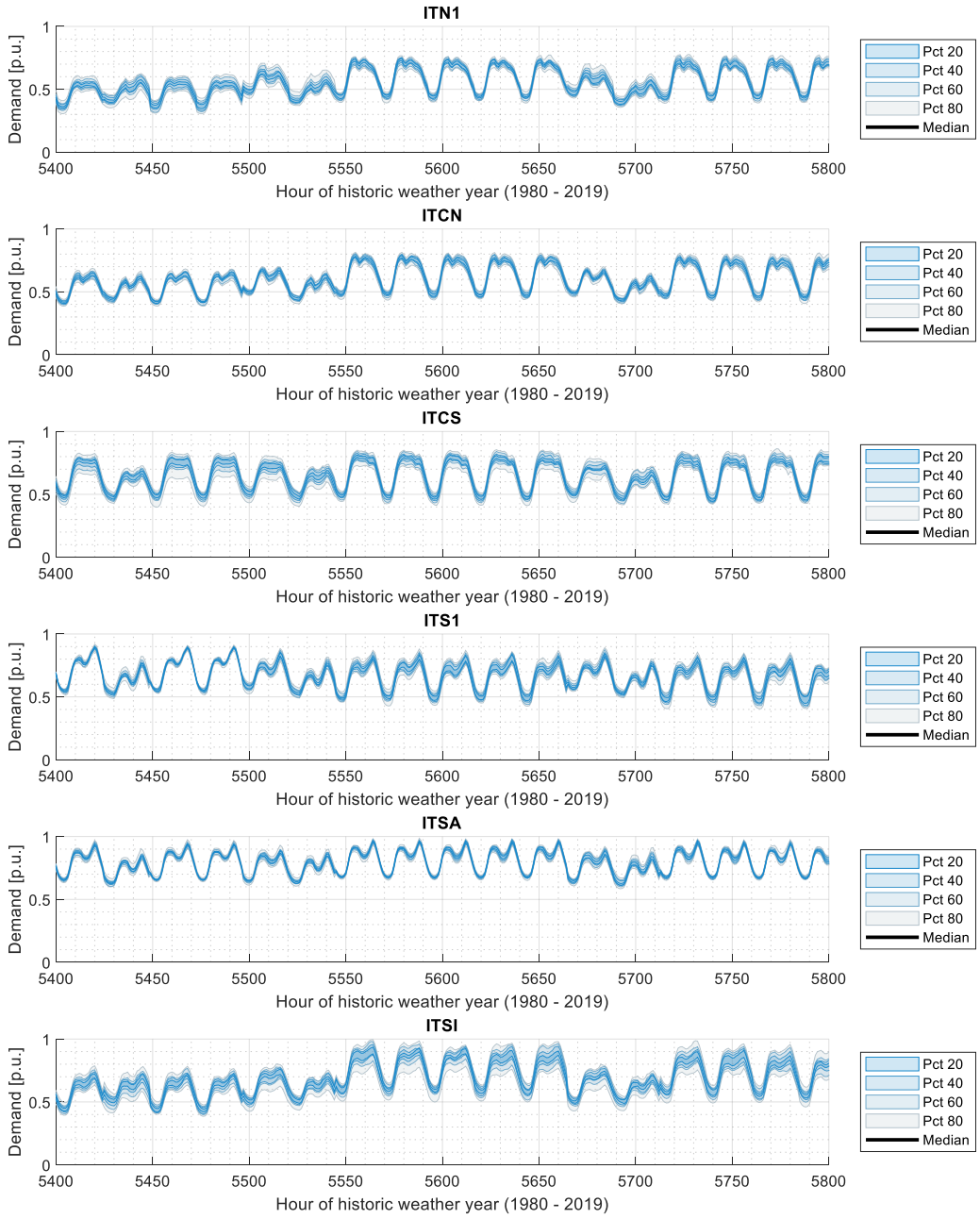

Figure 4-21: Exemplary visualization of load capacity factors for Italy's market areas (raw data taken from ENTSO-E [102])

The nodal load time series is generated in a similar fashion as in the case of hydro generation (Figure 4-17). Starting from nodal demand provided by MILES the scenario generation approach assigns each node to exactly one ENTSO-E market area. As a consequence, the corresponding time series of hourly load capacity factors for this market zone is then elementwise multiplied with the installed load of each node inside this area to generate demand time series. Thereby the installed load per node is defined as the reference demand time series' maximum value over time at each specific node.

Modelling spatial correlations

The time series of load differences with respect to the reference year per market zone are broken down on the nodal level. Therefore, it is assumed that the spatial distribution of deviations due to temperature is proportional to the hourly share of the nodal load with respect to the total load. To generate the final nodal load time series including uncertainty the reference time series coming from WP 4 and the calculated delta time series are summed up.

$$P^{\text{Load}}_{n,t,m,s} = P^{\text{Load,Ref}}_{n,t,m,s_{\text{Ref}}} + \frac{P^{\text{Load,Ref}}_{n,t,m,s_{\text{Ref}}}}{P^{\text{Load,Ref}}_{c,t,m,s_{\text{Ref}}}} \cdot \left(\Delta P^{\text{Load,Heat}}_{c,t,s,s_{\text{Ref}}} + \Delta P^{\text{Load,Cool}}_{c,t,s,s_{\text{Ref}}}\right) \forall t \in T$$

As there is no detailed information on the spatial distribution of load's temperature sensitivity, spatial correlations are modelled indirectly through the application of *MILES* by allocating the national demand to individual transmission grid nodes. As a consequence, using nodal weighting factors based on the hourly reference demand seems to be more realistic than an equal distribution of the differences over all grid nodes.

4.2.4. Outages of thermal power plants

The scenario generation and reduction approach developed in this task is dedicated to tackle long-term uncertainties. Short-term uncertainties in terms of grid related outages, e.g. the outage of transmission lines, can be considered in the new planning tool itself rather than in the input time series. However, mid-term uncertainties such as planned outages of thermal power plants can impact the operational scenarios results heavily. As such, thermal power plant availabilities have been integrated as a stochastic input to new planning tool. As detailed power plant information is collected to build the Pan-EU scenarios and market simulations are carried out to derive the border conditions for the regional cases, an existing method of the *MILES* simulation framework is applied to generate unit by unit availability profiles. For sake of completeness the modelling approach is presented briefly.

Thermal power plants are subject to periods of **planned outages** to refuel the units or perform preventive maintenance operations. These non-availabilities are typically planned on an annual basis considering the expected seasonality of the stochastic inputs, e.g vRES injections and demand. Thus, maintenance scheduling can be computed only once and considered as constant for each operational scenario. In contrast, **forced outages** of individual units occur by chance. Thus, forced outages due to technical problems or other random incidents have to be treated as an uncertain input in the planning tool.

Generating time series data

To consider power plant outages a **binary** time series is created where a units' availability is set to 1, when the plant is generally ready for operation and to 0, if the plant is not (i.e. it is in an outage state). The binary time series are generated for each thermal power plant based on two characteristic parameters [107]:

- Outage rate (r^{OUT}): Average proportion of time in which a unit is out of order.
- Outage duration (Δt^{OUT}): Typical duration until a unit is back in operation.

The characteristic parameters can differ significantly depending on the fuel-type and the age of a plant [107]. The failure probability (p^{Fail}) corresponding to the probability of a failure in $t+1$ assuming the unit was available in t_0 is defined in accordance [107] to as:

$$p^{\mathrm{Fail}} = \frac{r^{\mathrm{OUT}}}{r^{\mathrm{OUT}} + \Delta t^{\mathrm{OUT}} \cdot (1 - r^{\mathrm{OUT}})}$$

The algorithm generating daily outage time series for individual units is visualized in Figure 4-22

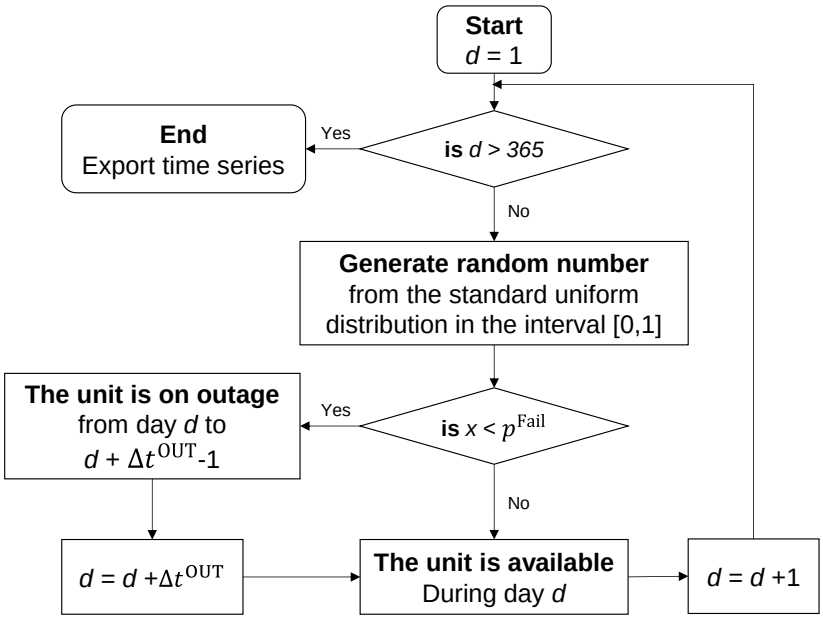

Figure 4-22: Algorithm to generate thermal power plant availabilities considering forced outages (adapted based on [67])

For each day *d*, a random number *x* [0, 1] is generated based on a standard uniform distribution. It is assumed that the plant is initially available on the first day (*d* = 1). If the random number *x* is larger than the failure rate, the unit stays available in the considered day *d* and the iterative procedure is repeated for next day. In case the randomly generated number is smaller than the failure rate, the unit is set unavailable for its specific outage duration. The iterative process continues from the day the unit becomes available again (d+$\Delta t^{\mathrm{OUT}} - 1$) until the end of the year. A more detailed discussion of the algorithm can be found in [107].

4.3. Outputs / Results

The methodology previous sections (cf. section 4.2.1 - 4.2.3) provides a number of time series for each transmission grid node considered in the planning model, taking into account generation and load capacity factors, of a maximum of 40 different realizations. As such, each operational scenario includes a set nodal time series data (8,760 time steps) for each Monte Carlo year sampled by historical climatic conditions from (1980 - 2019).

An operational scenario is formed by a set of nodal generation and load time series for the stochastic inputs of the planning problem namely vRES and hydro generation as well as the electrical demand. As depicted in Figure 4-23 the output of the scenario generation approach can graphically be interpreted as a set of three-dimensional matrices including the nodal generation and load data for each time step and realization.

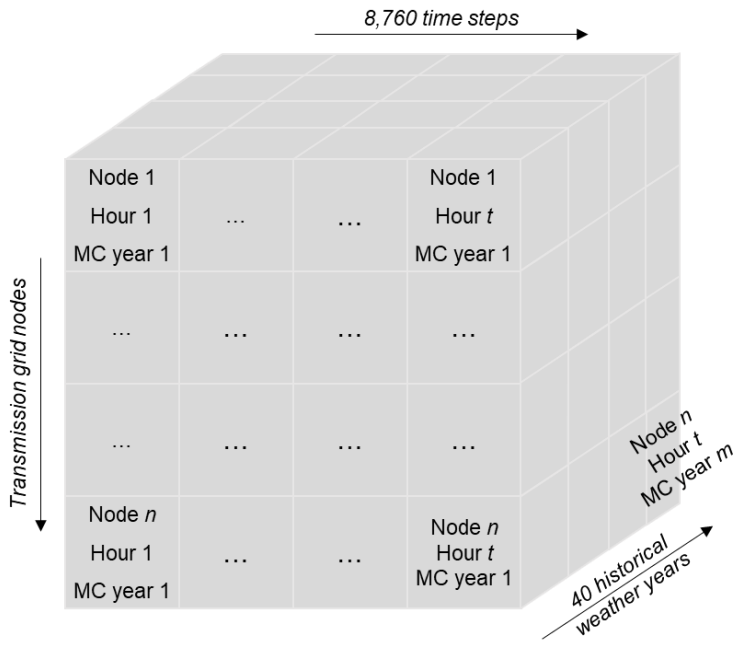

Figure 4-23: Schematic visualization of scenario generation's results

For each uncertain input parameter in terms of wind, solar and hydro generation as well as the load such kind of 3D matrix is generated including hourly feed-in and demand per node and realization. Additionally, hydro generation constraints in terms minimum and maximum production capacities, reservoir levels and inflows are generated per market zone.

Taking into account the grid model's level of detail (5,100 nodes for the entire Entso-e region), the hourly resolution of the scenario data (8,760 time steps), the amount of historical weather data (40 years) and the number of macro-scenarios (3) per target year (2030-2040-2050) as well as a variety of thermal generator outages, the number of input parameters generate becomes very extensive. Thus, a scenario reduction is applied to reduce complexity as described in the next section.

5. Scenario Reduction Approach

5.1. Planning input requirements and nomenclature

The input of the planning problem, to be provided by the scenario reduction, is the reference demand and generation. These are defined as (see section 2.1):

- Reference demand: $P^{ref}_{u,t,y}, \forall u \in S_u, \forall t \in S_t, \forall y \in S_y$
- Reference generation: $P^{ref}_{g,t,y}, \forall g \in S_g, \forall t \in S_t, \forall y \in S_y$

With S_u the set of demand elements, S_g the set of generators (Wind, PV, Hydro generation), S_t the set of time steps in the planning horizon, and S_y the set of planning target years. Within the FlexPlan project 2030-2040-2050 are defined as target years. It is assumed that generation and demand time series are required per node.

Within this document, one input *scenario* consists of a set of $P^{ref}_{u,t}$ and $P^{ref}_{g,t}$ values, i.e. the reference demand and generation time series of all load and generation (per network node) during 1 target year.

The objective of the scenario reduction method is to reduce the overall set of possible demand-generation scenarios, with size *m,* to a limited set of *s* scenarios, with a time series length of $n(S_t)$. A schematic overview of this procedure is shown in Figure 5-1.

In the scenario generation approach, see section 4, is explained that data on reference demand and generation will be provided per node for 40 climate years. The length of the time series required as input for the planning tool, $n(S_t)$, are initially set at one year of hourly time steps. Consequently, the overall space of possible planning scenarios, *m,* equals 40. The size of the reduced scenario set, *s*, depends on the computational performance of the planning tool, but is expected to be limited, at least to a number less than 10.

The computational performance of the planning tool needs to be sufficient to find a solution in a reasonable amount of time, therefore it may be required that the input time series are reduced in length (instead of a yearly time series, limited to e.g. one month, one day or one week). In that case, the generated scenarios will need to be split first into scenarios of the required length. Thus, if the time series of length $n(S_t)$ is set to one week by the planning tool computational time constraints, the 40 climate year scenarios have to be split up into 2080 (=40x52weeks) scenarios of 168 hourly time steps. As explained above, clustering will be used to achieve scenario reduction.

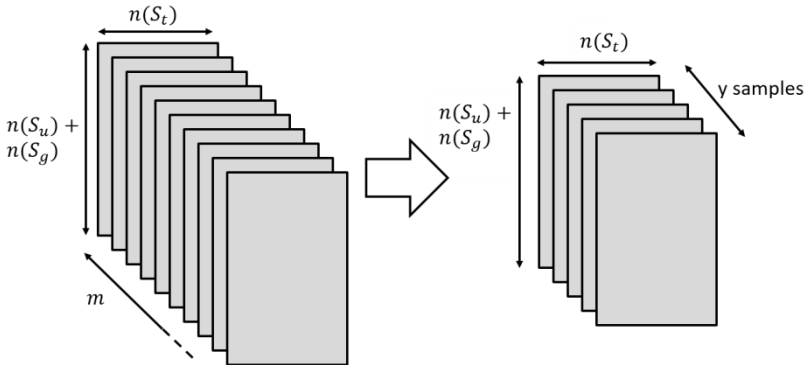

Figure 5-1 Schematic overview of the scenario reduction objective.

5.2. Scenario Reduction using clustering

Clustering is a machine learning technique that is used to find groups, i.e. clusters, of data objects within a dataset. Within the context of scenario reduction for transmission network expansion planning, the objective of the clustering is to find a limited, yet representative, set of input scenarios of the planning problem. A clustering algorithm detects clusters of 'similar scenarios' within the overall dataset of possible scenarios. It is then assumed that solving the planning problem for the set of input scenario's that consists of one scenario per cluster will lead to a solution that is the same as or close to the solution that would be calculated if the complete set of possible input scenarios would be used.

To use clustering, the data objects must have common features by which they are compared. The network state, or stated differently, the need for transmission capacity or flexibility, is defined by the load and/or renewable generation at each network node. Similar network states are thus defined by a similar combination of nodal load and generation. Both load and generation vary over time, as such, load/generation time series can be considered similar when a similar variation in the load/generation time series is observed. Within the network expansion planning problem, one data object consists of (all of) the time series of load and (renewable) generation at all of the nodes of the network. The data objects thus have features in two dimensions:

- Node dimension: the value of load/generation at each separate node, or the power of each demand/generation element can be considered as separate features. The number of node dimension features equals $n(S_g) + n(S_u)$.
- Time dimension: the node values at each separate time step of the time series can be considered as separate features. The number of time dimension features equals $n(S_t)$.

In total, one observation or one data object thus reaches $(n(S_g) + n(S_u)) \cdot n(S_t)$ features. Given the size of the considered networks, as well as a minimal required time series length, this number of features become very large.

It must be noted that using observations with a very large number of features holds the risk of suffering from the so called 'curse of dimensionality' [90]. In this case, feature reduction techniques should be used to make sure sensible clusters are produced from the dataset.

5.2.1. K-means clustering for scenario reduction

A plethora of different clustering methods exist in the literature. These algorithms are often categorized as partitional clustering and hierarchical clustering methods [91].

Hierarchical clustering methods construct the clusters by recursively partitioning the data objects. In a bottom-up approach, where each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained. The result of the hierarchical methods is a so-called dendrogram. A dendrogram is a tree-shaped diagram which represents the nested grouping of objects (i.e. when branches of the tree are joined) and similarity levels at which groupings change. A clustering of the data objects is then obtained by cutting the dendrogram at the desired similarity level. The name "hierarchical clustering" comes from the result of the clustering which gives an extensive hierarchy of clusters that merge with each other at certain distances.

A partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique. Partitional methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. These clusters should fulfil the following requirements: each cluster must contain at least one data object, and each object must belong to exactly one cluster. One of the most popular partitional clustering algorithms is K-means clustering.

K-means clustering finds the k cluster centres, also referred to as centroids, and assigns the objects to the nearest cluster centre such that the squared distances from the cluster are minimized [78]. When using K-means clustering, a predefined number of clusters needs to be set, as well as the distance metric used. The problem that is solved with K-means clustering is the following:

$$\arg\min_{K} \sum_{i=1}^{k} \sum_{x \in K_i} \|x - \mu_i\|^2$$

With $x = (x_1, x_{2,\ldots,} x_n)$ the original set of objects (i.e. within the FlexPlan project, the original set of scenarios), each forming a multidimensional vector, and with $K = (K_1, K_{2,\ldots,} K_k)$ the resulting clusters, where μ_i is the mean of set K_i (i.e. the cluster centroid). This optimization problem is solved iteratively. A number of variants of this algorithm exists, e.g. where not the mean but the median is used (k-medians), or where the centroid is restricted to be a member of the cluster (k-medoids).

In the literature, it has been shown that the number of clusters needs to be high enough to be able to cover the full space of operating conditions with sufficient accuracy [69] [72]. One drawback of K-means clustering is that it may converge to a local optimum, depending on its starting condition (i.e. choosing the initial centroids), therefore it is often advised to run the algorithm several times, with different initialization conditions. Because of the widespread knowledge and application of K-means as clustering method, this method will be used for scenario reduction in this project. Also, K-means clustering is known to be able to handle large datasets [92]. However, it must be noted that other clustering methods could be applied within the scenario reduction context.

To determine the similarity/difference between the objects in the given dataset, a similarity metric needs to be defined. For example, if the objects within a dataset have vectors of coordinates as features, a similarity measure can be interpreted as the distance (length, vector norm) between the data objects. The smaller the distance gets, the closer (i.e. the more similar) the data objects are. An often -used metric to define the similarity of load/generation time series is the Euclidian distance, e.g. used in [85] and [75], and will therefore be also used within this project. The Euclidian distance d between two data objects x_i and x_j with features f is defined as:

$$d(x_i, x_j) = \sqrt{\sum_f \left(x_{i,f} - x_{j,f}\right)^2}$$

A drawback of using the Euclidian distance as a similarity metric is the tendency of the largest-scaled feature to dominate the others. In the context of load/generation time series, this would entail that the variation of the largest generator (or load) within the system considered would dominate the choice of clusters, leading to a possible underrepresentation of other important events in the system. The solution to this problem is to normalize the features, i.e. the power values, to a common range. Therefore, it is proposed

to normalize the load and generation power values to their nominal capacity, which leads to feature values to be between 0 and 100%.

To illustrate the clustering approach, Figure 5-2 shows the result of the clustering of weekly wind profiles using K-means clustering. In this example, 40 years of hourly wind time series are split up in weekly profiles (2080 weekly profiles), and then clustered into ten clusters. Each subplot shows the weekly profiles of 1 cluster, the centroid of the cluster is shown as a black line. As stated above, the centroid is defined as the 'multidimensional mean' of the cluster, the centroid is (not necessarily) a member of the original set.

To illustrate hierarchical clustering, Figure 5-3 shows the result of the clustering of weekly wind profiles using Ward's hierarchical clustering algorithm. The dendrogram shows that at a cutoff distance of 15, 10 clusters can be identified. These clusters are separately plotted in Figure 5-4. In this figure, each subplot shows the weekly profiles of 1 cluster. As can be noticed, the result of the hierarchical clustering differs slightly from the clustering resulting from K-means clustering.

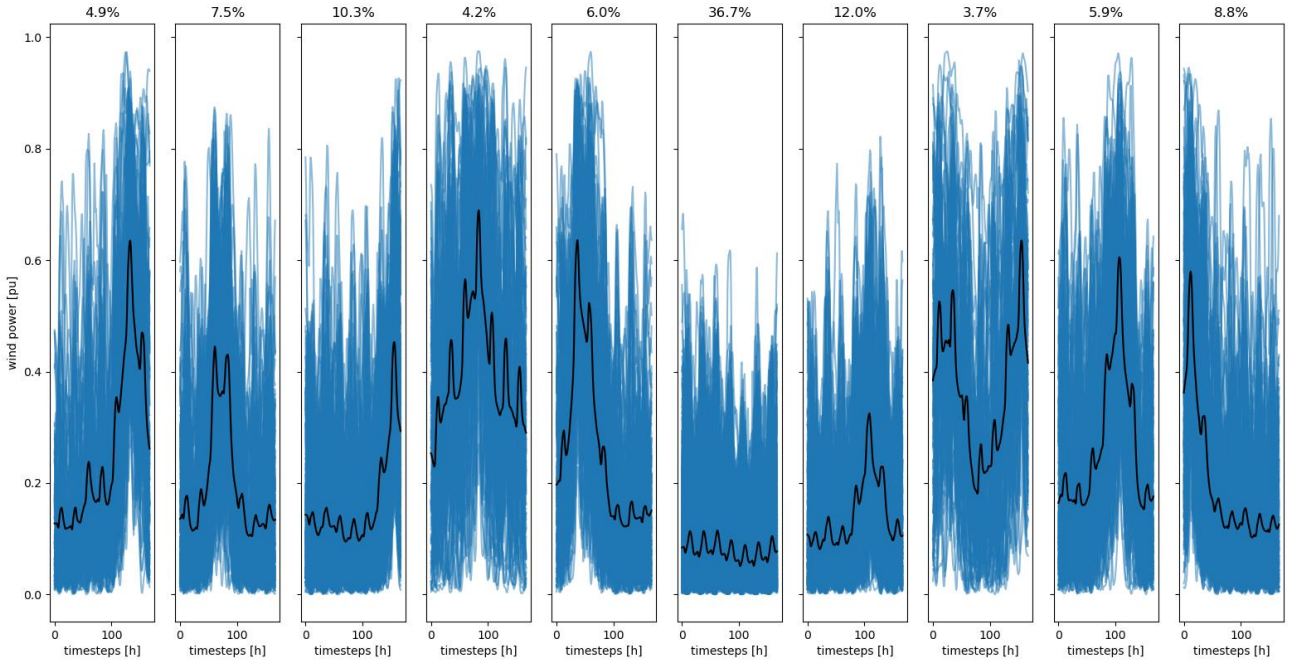

Figure 5-2: Illustration of K-means clustering. Weekly wind power profiles are clustered in 10 clusters. Each subplot shows all profiles of 1 cluster, the black lines are the centroids of each cluster. The probability of each cluster is indicated above each subplot.

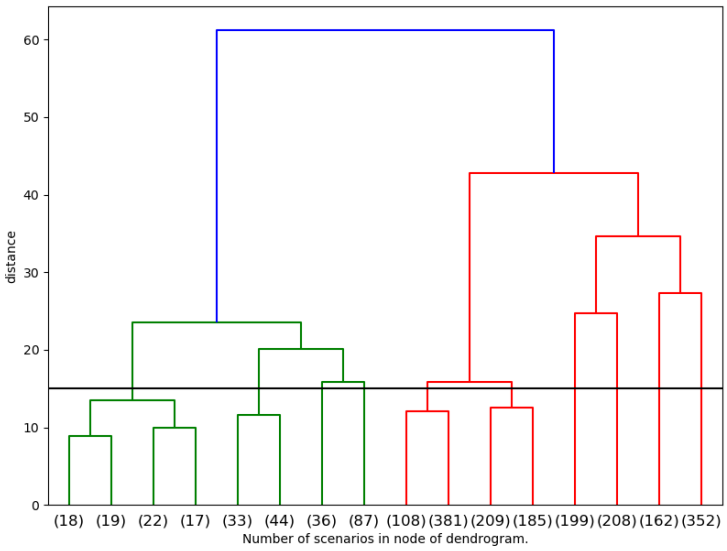

Figure 5-3: Illustration of hierarchical clustering. Weekly wind power profiles are clustered using Ward's hierarchical clustering, the figure shows the resulting dendrogram. At a cut-off distance of 15, indicated by the black horizontal line, 10 clusters are obtained.

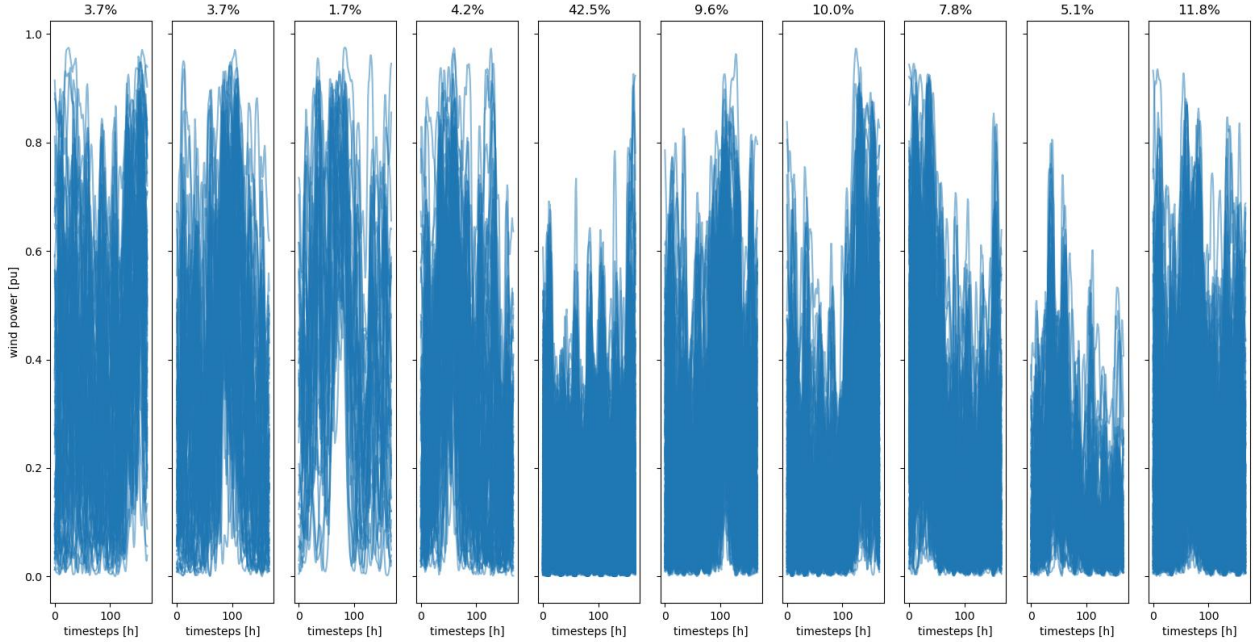

Figure 5-4: Illustration of hierarchical clustering: Weekly wind power profiles are clustered using Ward's hierarchical clustering, resulting in 10 clusters. . Each subplot shows all profiles of 1 cluster, the probability of each cluster is indicated above each subplot.

5.2.2. Feature reduction

As mentioned, when using observations with a very large number of features, clustering holds the risk of suffering from the so called 'curse of dimensionality'. This means that too many dimensions cause every observation in the dataset to appear equidistant from all the others. In these cases feature reduction techniques should be used to make sure that sensible clusters are produced from the dataset.

The considered data objects have $(n(S_g) + n(S_u)) \cdot n(S_t)$ features. When the considered network size, as well as the required time series length are large, feature reduction techniques might become necessary.

Many feature reduction techniques and algorithms have been developed in the literature. Within this project, three approaches are proposed for possible application, namely Principal Component Analysis (PCA) and two variants of feature selection. **The choice of the feature reduction technique strongly depends** on the actual datasets and network characteristics.

Feature reduction with PCA

Principal component analysis is very often applied for feature reduction [94] of time series data. PCA is defined as an orthogonal *linear* transformation that transforms the data to a new coordinate system while preserving as much 'variability' (i.e. statistical information) as possible. This means that 'preserving as much variability as possible' translates into finding new variables that are linear functions of those in the original dataset, that successively maximize the variance and that are uncorrelated with each other. Finding such new variables, i.e. the principal components, reduces to solving an eigenvalue/eigenvector problem. The number of components to keep to retain maximal variance from the original dataset depends on the dataset used. The details of the techniques to be used can be found in literature, an extensive overview is given in [94].

The downside of PCA, and by extension of all feature reduction approaches that rely on applying a transformation of the dimension space, are that the physical meaning of the obtained components is often lost. Another attention point is that it is advised that the datasets on which PCA is applied, need to have a sample size that is (much) larger than the number of features, for the method to perform well [109].

Within the FlexPlan project, PCA can be applied to the input datasets when the use of clustering on the original datasets shows to result in non-optimal clusters, because of the high number of dimensions in the dataset. To apply PCA, it is proposed to rely on existing implementations that are available as open-source libraries, e.g. the scikit-learn package for Python [93]. An illustration of the technique is given below, in section 6.2.

Feature selection

As an alternative approach for feature reduction, next to 'transforming' the features to a lower-dimensional parameter space, it is also possible to reduce the features by selecting the most significant features. This approach has the benefit that the physical meaning of the data objects is kept.

Defining significant scenario features

In first feature selection approach, a number of significant characteristics of the scenarios can be defined. A non-exhaustive list of significant characteristics that define, or specify, the different scenarios within the scenario set may be:

- Total demand, summed over all nodes and time steps
- Total generated wind, PV and hydro power
- Maximal demand
- Maximal generated wind, PV and hydro power
- Minimal demand,
- Minimal generated wind, PV and hydro power
- Maximal demand variation, i.e. maximal overall demand difference between two time steps
- Maximal PV, wind, hydro variation
- …

To perform feature reduction by this form of feature selection, the chosen significant features have to be computed for each scenario within the original scenario set. Then, scenarios are represented by their significant features only. Clustering can then be performed on the set represented by its reduced features only (after normalization).

The upside of this approach is that the physical features of the scenarios are kept, but at a minimal level however. The downside is that it is not straightforward to define a small set of significant features where the nodal variation of demand and/or generation within the feature set is still visible.

Clustering the node dimension

As a third method to reduce the number of scenario features, it is proposed to select only the time-dimension features and reduce the node dimension by applying clustering on the node-dimension features.

With a dataset consisting of time series of t time steps and m variations of time series for each network node, the overall number of network states becomes $t \cdot m$. A network state consists of a vector of n power values for a network with n nodes, and thus each time step a different network state occurs for each load-generation scenario. These network-state vectors can be clustered according to their nodal power. The result of this

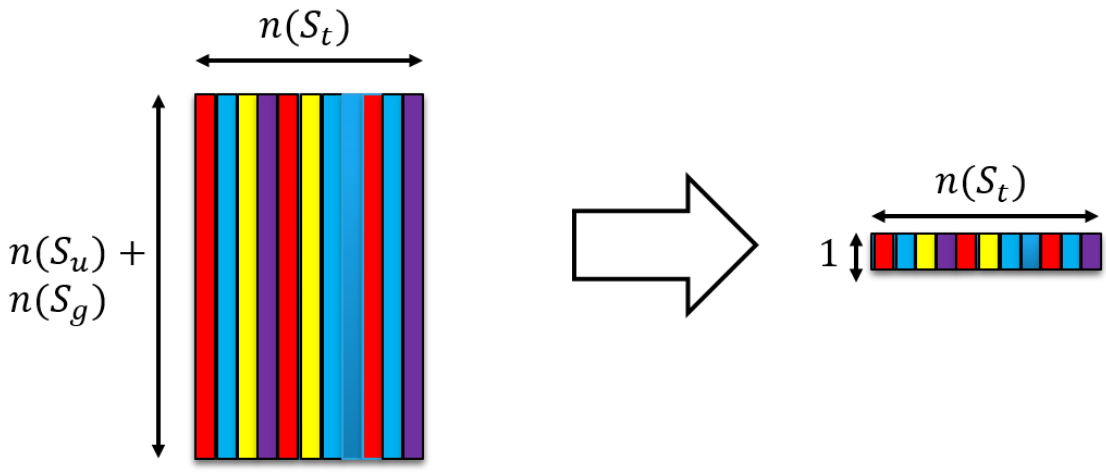

Figure 5-5: Feature reduction: reducing the node-dimension. Similar network states are indicated with the same colour. The feature reduction step results in a time series of cluster number values.

clustering step is m time series (of length t) of "*cluster number values*". Thus, for each load-generation scenario consisting of n time series of node power values, one time series of cluster number values is derived. The

$(n(S_g) + n(S_u)) \cdot n(S_t)$ features of the original dataset are thus reduced to a dataset of $n(S_t)$ features. This process is illustrated in Figure 5-5.

To be able to cluster the resulting lower-dimensional dataset to finally achieve the required reduced scenario set, the distance between the lower dimensional data objects has to reflect the distance or similarity between the objects of the original dataset. As mentioned previously in section 5.2.1, to compute the similarity between the scenarios, the Euclidian distance will be used. For network expansion planning problems, it is expected that the amount of renewable (non-dispatchable) generation produced at each time step is important to consider, as well as the amount of nodal demand. Additionally, time series with a similar global network load can be very different for the planning problem in the sense that the load might be divided over the network nodes in a very different way. Analogously, this holds as well for the generation.

To make sure distance metrics are computed correctly, the cluster number values, of which the reduced dataset consists of, are ordered twice:

1. The cluster number values are ordered according to the overall non-dispatchable generation of their cluster centroids. Centroids may have a similar (or same) global generation (i.e. when the generation is distributed over different nodes), then those centroids are ordered according to the injection per node. Thus, time series with network states with a very different global renewable generation will be very different. Clusters with a similar global renewable generation will differ less, but they can still significantly differ if the generation is divided over the different nodes in a very different way.
2. Analogously, the cluster number values are ordered according to the overall demand.

In this case, the scenario reduction clustering step (as explained in section 5.2.1) has to be applied twice, once for each value of cluster numbering above. The original data objects are then part of two groups. The first group is based on its renewable generation, and the second group based on its demand. The number of theoretically possible group-combinations is limited and depends on the size of the clusters used in the scenario reduction clustering step.

To reach the final scenario clustering result, all scenario data objects with the same renewable-based and demand-based group number are grouped together.

The advantage of this feature reduction method is that the effect of different nodal values remains in the reduced dataset. The disadvantage is that the method is more complex, and when scenarios having many time steps need to be reduced, the number of features can still be too large. An illustration of this method is given in section 6.2.

5.3. Overall scenario reduction methodology

The dataset to be clustered, consists of 40 year-long time series of hourly time steps (i.e. 8,760 time steps) for each network node. The result of the clustering are y time series (of length $n(S_t)$, with $n(S_t)$ equal to 8,760 or less) of load and generation power values. These y time series are chosen as the centroids of each cluster after K-means clustering is applied to the overall dataset. The size of each cluster is an indication of the probability of occurrence of that combination of load and generation present in that cluster. These probabilities can also be provided as input to the planning optimization tool, e.g. if a stochastic planning approach is envisaged. Depending on the computational performance of the planning tool, a feature reduction step needs to be applied to the overall dataset before the clustering step, to make sure that sensible results are produced.

To conclude, the overall clustering-based scenario reduction methodology is outlined as follows:

1. (if required) Split the yearly scenarios into scenarios with a (time-dimension) length equal to the setting required by the planning tool. The length is set to 24 time steps if representative daily scenarios are required, to 168 time steps, if weekly scenarios are required, etc.
2. Normalize the load and generation time series as produced by the scenario generation, by the nominal capacity of the respective loads and generators.
3. (If required) apply feature reduction by (a) applying PCA, (b) selecting and computing significant features, or (c) reducing the node dimension by clustering the node-dimension features.
4. Perform K-means clustering on the overall, but possibly feature-reduced, dataset. The requested number of clusters *s* is defined by the computational efficiency of the planning tool and is provided as an input to the scenario reduction.
5. Choose one representative scenario for each cluster. The computed cluster centroid is an option, but also a randomly chosen scenario belonging to that cluster can be used. If feature reduction was applied, choosing the cluster centroid is not an option, as the centroid in this case will not have the same dimensions as the original scenarios.
6. Rescale the chosen scenarios back to their original values.
7. (if required) Identify the importance of the representative scenario's according to their cluster size.

6. Numerical Example

To validate the scenario generation and reduction methodology with regard to its feasibility and computational performance a test case has been defined. A modified version of the 6-Bus Garver test system [108] applied to Italy, as defined for other testing activities within the FlexPlan project, is used to test the scenario generation and reduction method. Each node is located in one of the Italian market areas, as can be

Figure 6-1: Geographic overview of the 6-Bus test case

seen in Figure 6-1.

The defined geographic locations of the six nodes are given in Table 6-1. Two PV generators have been placed in the system, replacing conventional generators. The first generator is installed at node 3 (IT-CS) with a capacity of 180 MW. The second one has a capacity of 240 MW and is located at node 6 (IT-SAD). Furthermore, there is one wind generator with a capacity of 240 MW at node 6 (IT-SAD). The assumed annual peak load at each location is also given in Table 6-1.

Table 6-1: Node names, geographical location, and load and renewable generation properties of the test case.

Node	Node name	Longitude	Latitude	Annual peak load [MW]	PV generation [MW]	Wind generation [MW]
1	Italy central north	43.4894°	11.7946°	80	/	/
2	Italy north	45.3411°	9.9489°	240	/	/
3	Italy central south	41.8218°	13.8302°	40	180	/
4	Italy south	40.5228°	16.2155°	160	/	/
5	Sardinia	40.1717°	9.0738°	240	/	/
6	Sicily	37.4844°	14.1568°	/	240	240

6.1. Scenario generation

For the proposed example network, scenario data was generated using the method outlined in section 4. Wind, PV and demand profiles were generated for each network node for 40 climate years. Figure 6-2 shows the first 1000 time steps of the wind and PV generation time series of all 40 generated scenarios. In this figure the sum of the wind, PV and demand profiles are plotted.

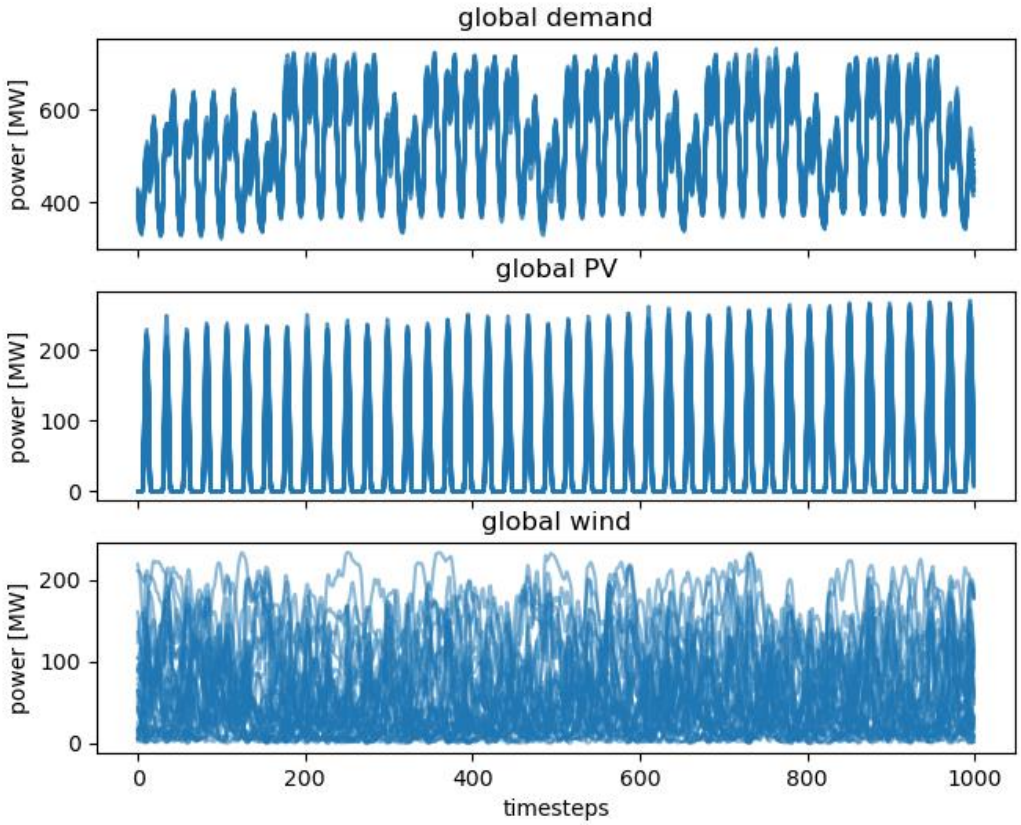

Figure 6-2: A plot of the 40 generated scenario's first 1000 time steps of the global demand, pv and wind profiles of the network

In Figure 6-3 the first 1000 time steps of the nodal balance (i.e. nodal demand minus nodal generation) of the 40 generated scenarios are shown.

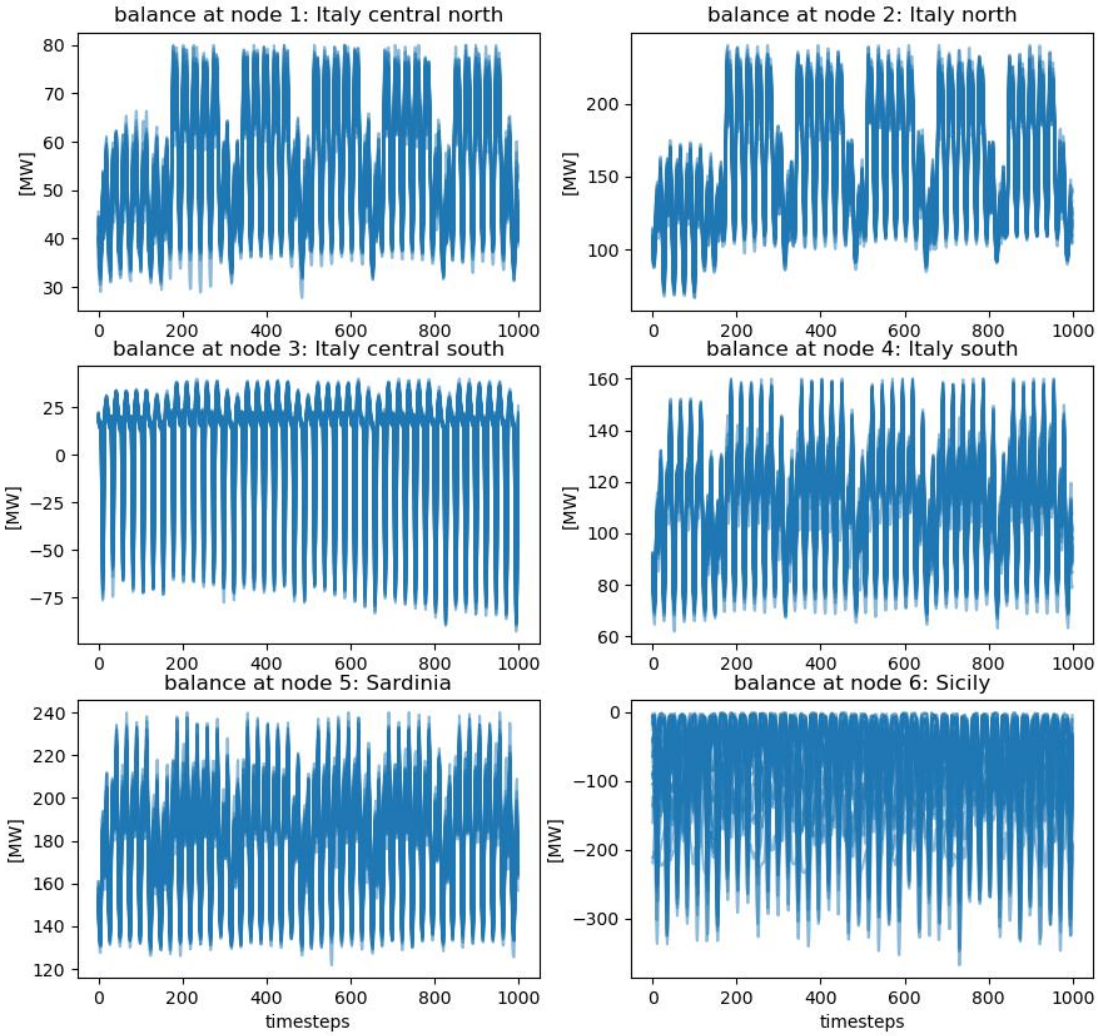

Figure 6-3: A plot of the 40 generated scenarios: the first 1000 time steps of the node balance of each network node.

6.2. Scenario reduction

The purpose of the scenario reduction is to choose a representative but reduced set of time series out of the 40 generated scenarios. Of the 40 generated years, there are 35 years with data on demand and PV as well as wind, all correlated on time. Those 35 correlated years are used as input for the scenario reduction.

To illustrate the scenario reduction methods, two sets of reduced scenarios were created for the example, a first set with time series length of 8760 time steps (i.e. one year), and a second set of with a time series length of 24 h (i.e. one day). For each of the sets, different feature reduction methods were applied, and the different clustering results compared in the following paragraphs.

In a follow-up step of the work presented in this deliverable, the impact of the clustering and specific feature reduction methods on the outcome of the planning optimization tool will be examined as part of the broader testing activities of the project. Within those tests, an assessment will be made on how many clusters are required to achieve a good balance between obtaining reliable planning tool results and having acceptable computation times. The same evaluation will be done to determine the impact of shorter input scenario time series on the planning tool decision making.

To create the clusters for this example, the clustering algorithms available in the Python package Scikit-Learn [93] have been used.

6.2.1. Scenario reduction to yearly profiles

In a first scenario reduction example, the size of the reduced set of scenarios is (arbitrarily) chosen to be 5, and the length of the representative scenarios is one year. The result of clustering the 35 scenarios into 5 clusters is shown in Figure 6-4. In Figure 6-4, and following similar plots, each column indicates one cluster, and each row shows the demand, PV and wind profiles of all scenarios belonging to that cluster, in per unit

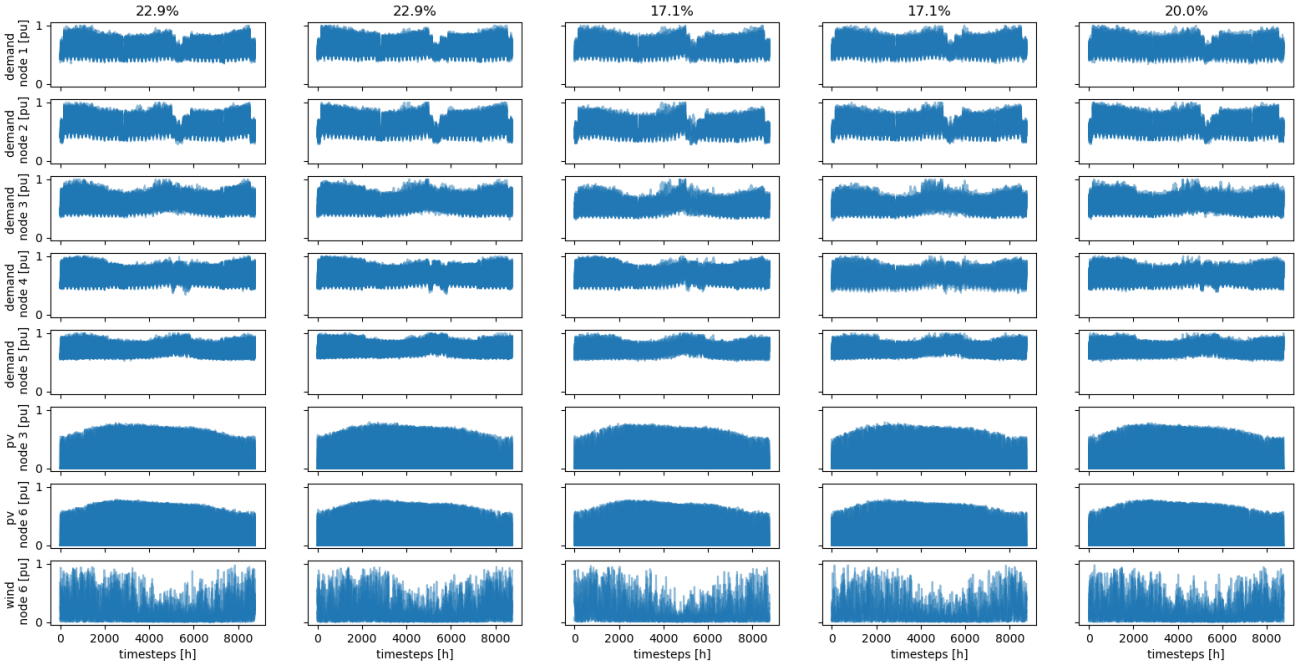

Figure 6-4: Result of clustering the 35 climate-year scenarios. Each column shows the demand, PV and wind profiles of all scenarios in one cluster. Above each column the probability of each cluster is indicated.

value. Above each column the probability of each cluster is indicated. As can be seen from the plot, the scenarios differ mostly in terms of wind power. The yearly demand and PV for each cluster are rather similar.

Figure 6-5 shows the global load, PV and wind profiles of the cluster centroids (left column in the figure) and one representative scenario of each cluster, chosen as a random pick from each cluster (right column in the figure). For clarity reasons, only the first 1000 time steps of each scenario are shown. The cluster centroids have less 'peaks' in their profiles, compared with the random chosen scenarios. A lack of power peaks in the scenarios might be a reason not to use cluster centroids as input for a planning tool, as non-optimal decisions might be taken.

The pairwise distances between the 35 scenarios, ordered according to cluster number, are shown in Figure 6-6. As mentioned in section 5.2.1, the Euclidean distance metric is used to compute similarity between the different scenarios. As shown in the plot, the distances between all yearly scenarios do not differ significantly, leaving it difficult to filter out different clusters from the scenario set. The reason for the small differences is most probably due to the very large feature size, therefore feature reduction will need to be applied.

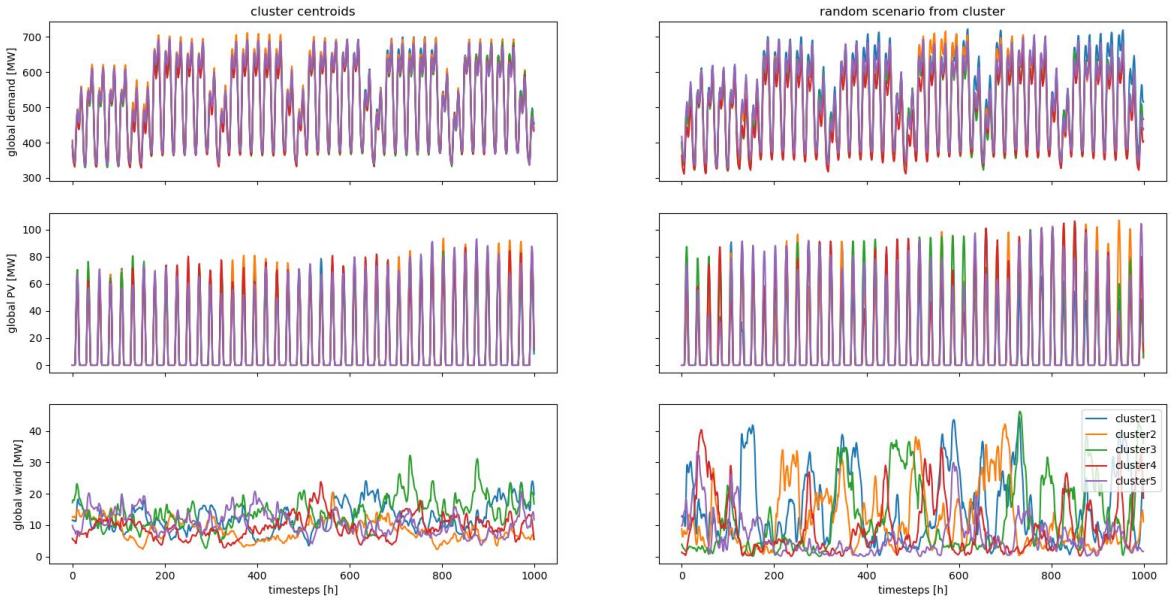

Figure 6-5: Global load, wind and PV profiles (first 1000 time steps are plotted), of the cluster centroids (left column) and one representative scenario of each cluster (right column).

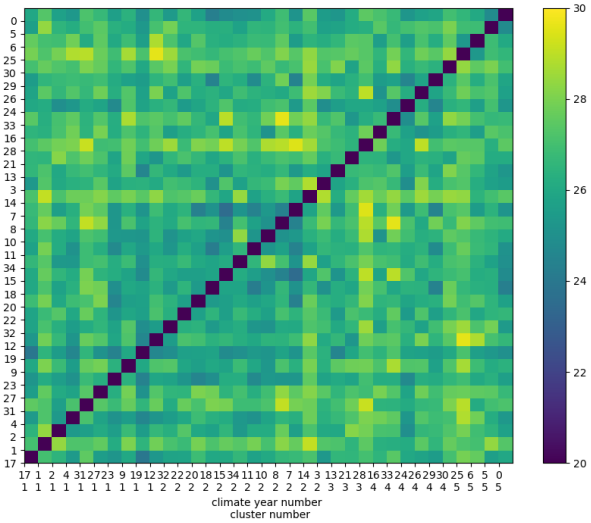

Figure 6-6: Pairwise Euclidian distance between generated scenarios: the distance between each scenario-pair is indicated as a colour. A larger distance is indicated by a lighter colour. The scenario number is indicated on the horizontal and vertical axes. The cluster number is also given in the horizontal axes.

To achieve an improved clustering result, PCA was applied to reduce the features of the scenarios. However, due to the very small example network, and very similar scenarios, clustering after feature reduction with PCA[2] resulted in rather similar clusters (but with different probabilities), as shown in Figure 6-7. In fact, it is advised to apply PCA as a method when the number of samples is (much) larger than the number of features, which is not the case for the example dataset., therefore, PCA is not the go-to method for feature reduction when year-long time series are required.

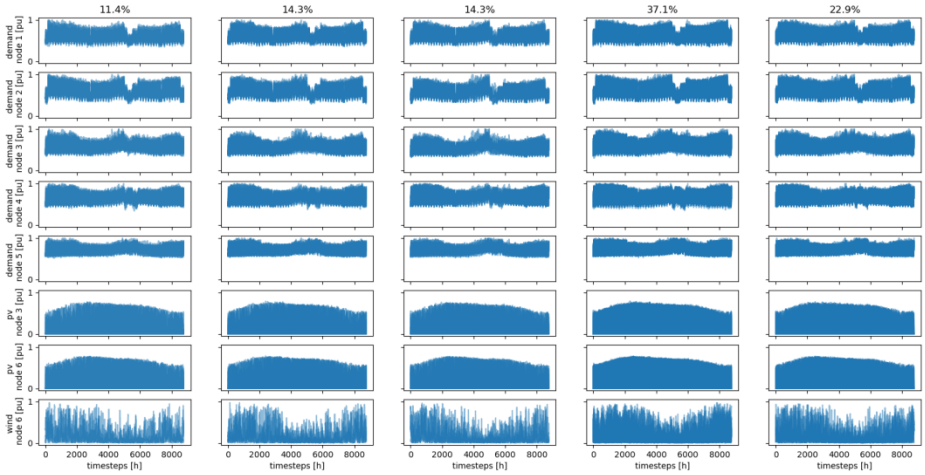

Figure 6-7: Result of clustering the 35 climate-year scenarios, after applying PCA. Each column shows the demand, PV and wind profiles per cluster. Above each column the probability of each cluster is indicated. Similar clusters are obtained as with the case without PCA.

[2] The number of components to keep was set to 30 in the PCA application. Analysis of the dataset showed that with this number, the variance retained in the dataset was above a threshold of 95%.

Additionally, feature reduction towards scenario characteristics has been applied. To achieve this, each scenario was characterised by 12 features, namely the features indicated in section 5.2.2. Those values are shown in Figure 6-8. In this figure, the characteristics are plotted for each scenario, and are coloured according to the cluster, to which the scenario belongs to.

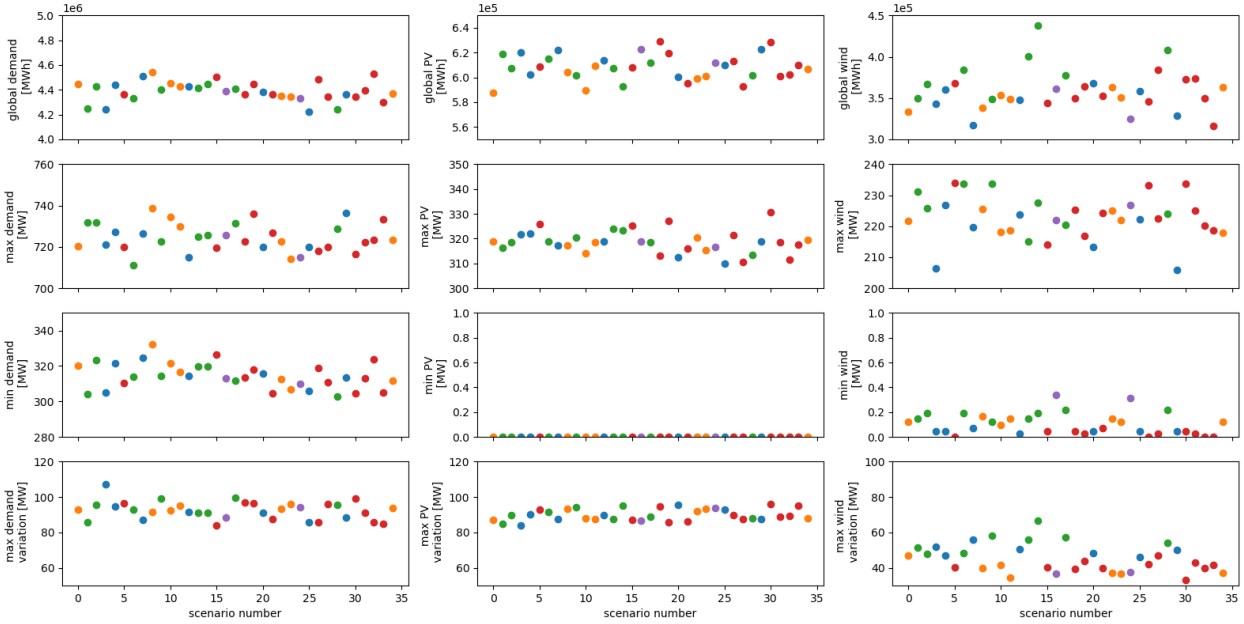

Figure 6-8: The computed characteristics of each of the 35 scenarios. Each subplot shows one characteristic for each scenario, coloured by cluster number.

The result of the clustering is shown in Figure 6-9. When comparing this clustering result with previous plots, it can be noticed that slightly different clustering results are obtained. When looking at the clustering results, in combination with the computed characteristics of each scenario, it seems that the produced results seem more relevant as input for the planning tool optimisation.

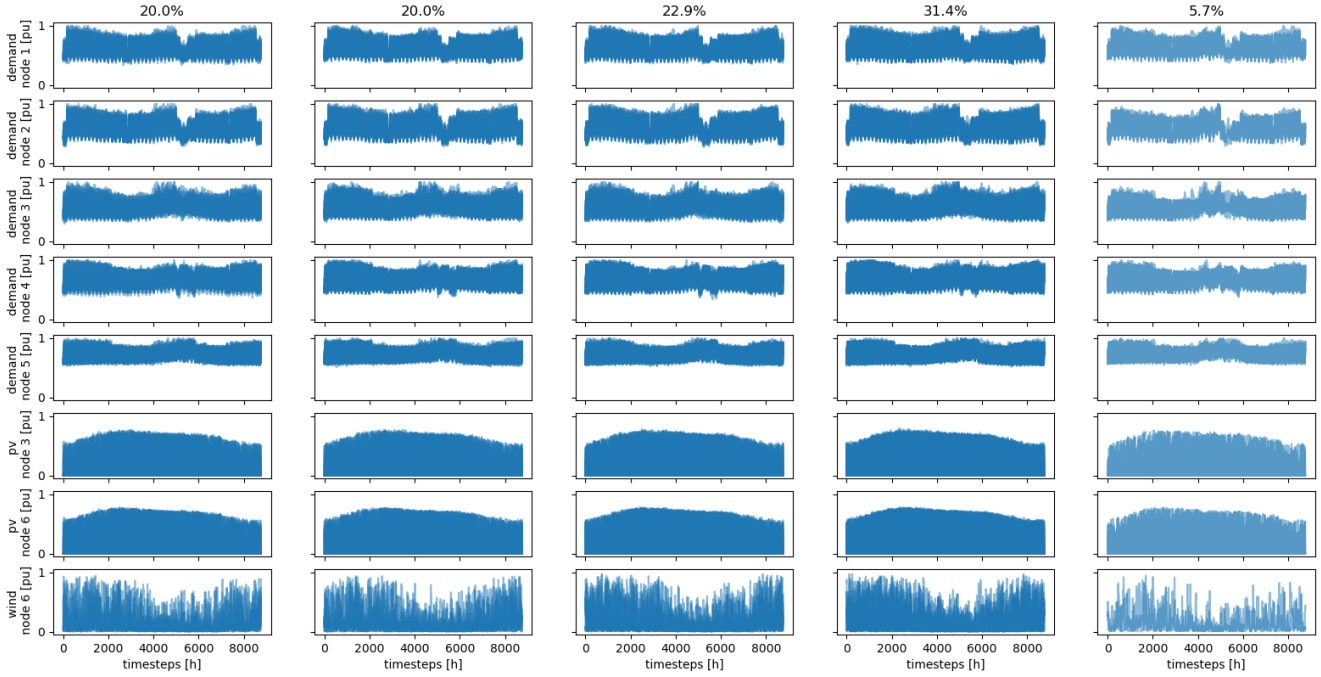

Figure 6-9: Result of clustering the 40 climate-year scenarios, after reducing the yearly time series to 12 characteristics.. Each column shows the demand, PV and wind profiles per cluster. Above each column the probability of each cluster is indicated.

Finally, also feature reduction by applying clustering on the node dimension was also performed on the test data. In this example, it is chosen to limit the number of possible grid states to 30. The resulting 30 'node dimension clusters' are visualised in Figure 6-10. This figure shows the centroids of each 'grid state cluster'.

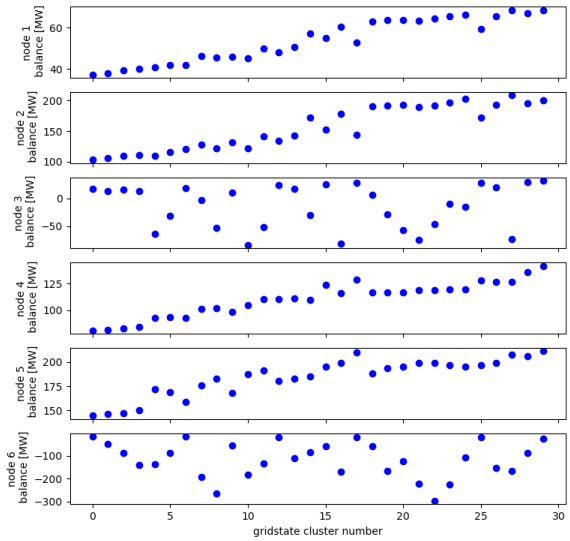

Figure 6-10: The identified possible grid state in the dataset: the grid states are reduced down to 30 possibilities.

Then, as explained in section 5.2.2, the scenarios, expressed as time series of 'grid state cluster numbers' are clustered in a subsequent step. The result of this clustering step is shown in Figure 6-11, and again different clusters are obtained. The obtained clusters are however quite similar to the ones obtained without any feature reduction applied.

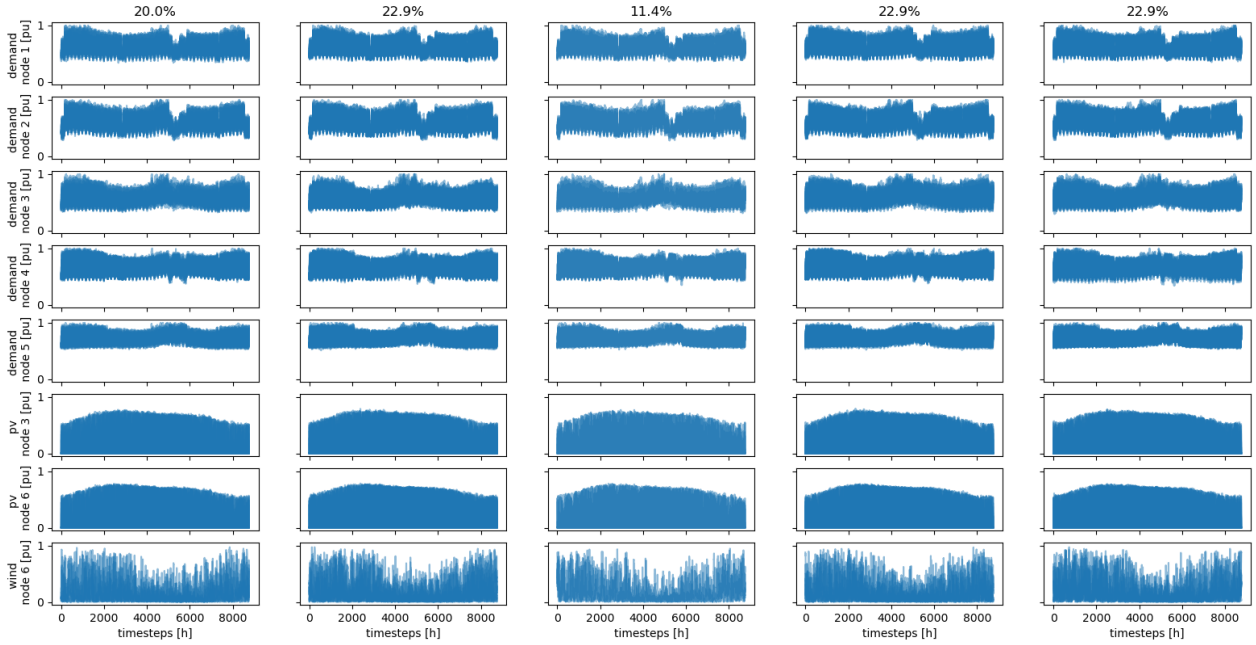

Figure 6-11: Result of clustering the 40 climate-year scenarios, reducing the scenarios to time series of grid states. Each column shows the demand, PV and wind profiles per cluster. Above each column the probability of each cluster is indicated.

From the scenario reduction results it can be concluded that for the 6-node example, the yearly differences within the original scenario set are not very large. Moreover, because of the size of the feature set, versus the number of samples in the original set, the clustering methods have difficulties finding sensible clusters within the original set. Also, applying PCA as feature reduction method, as well as reducing the number of features by applying clustering on the node dimension does not provide satisfying results. Therefore the approach where the yearly scenarios are clustered according to a limited set of 'main characteristics' will be followed when assessing the planning tool decision taking versus the number of scenario inputs, when those scenarios have a timeseries length of one year.

However, to determine which feature reduction method produces the most significant scenarios for the planning optimisation tool remains an open question. This cannot be determined based on the clustering results alone. Therefore within the planning tool testing activities, the planning tool decisions when fed with different scenario inputs will be compared to determine which feature reduction method performs best.

6.2.2. Scenario reduction to daily profiles

In a second scenario reduction example, it is assumed that the planning tool restricts scenarios to a length of one day (24 hourly time steps), because of computational reasons. The clustering exercise is now repeated to result in 5 representative daily clusters. In order to perform the clustering, the yearly scenarios are first split up in daily scenarios before clustering, resulting in a dataset size of 12775 samples, and each sample a size of 144 features (6 nodes times 24 time steps).

The resulting clusters after applying K-Means clustering without reducing the features on the dataset, are shown in Figure 6-12, with the probabilities of each cluster indicated on top of each column. It is clear that the different daily clusters are more distinct than the yearly clusters. For example, week and weekend-days are discovered by the clustering, visible in the demand curves.

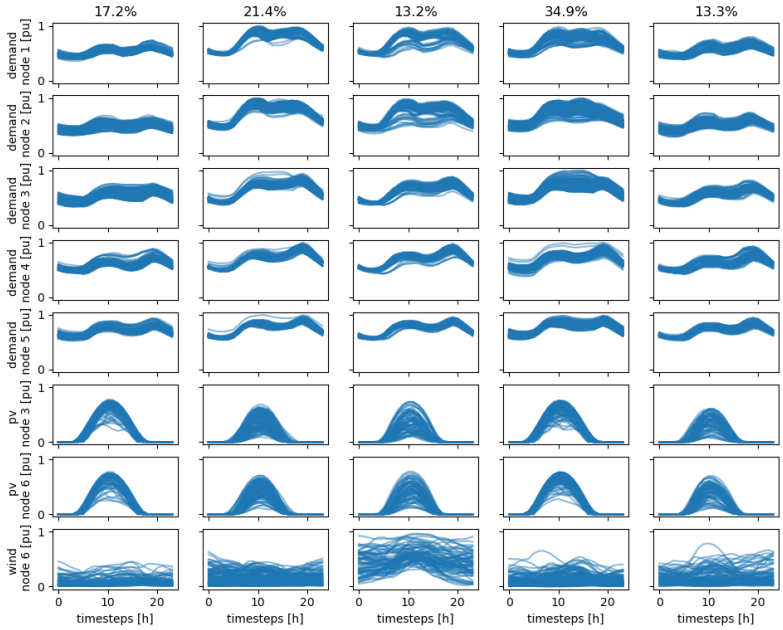

Figure 6-12: Result of clustering the 40 climate-year scenario's into daily profiles, without feature reduction applied. Each column shows the demand, pv and wind profiles per cluster. Above each column the probability of each cluster is indicated.

To investigate the influence of feature reduction on the daily-scenario clustering, each daily scenario was reduced to its 'most significant' characteristics (i.e. the same characteristics outlined in section 5.2.2). Figure 6-13 shows the clustering result of the feature-reduced dataset. The produced clusters are similar to the clusters shown in Figure 6-12. It seems that the reduced features reflect reasonably well the differences between the different days in the original dataset.

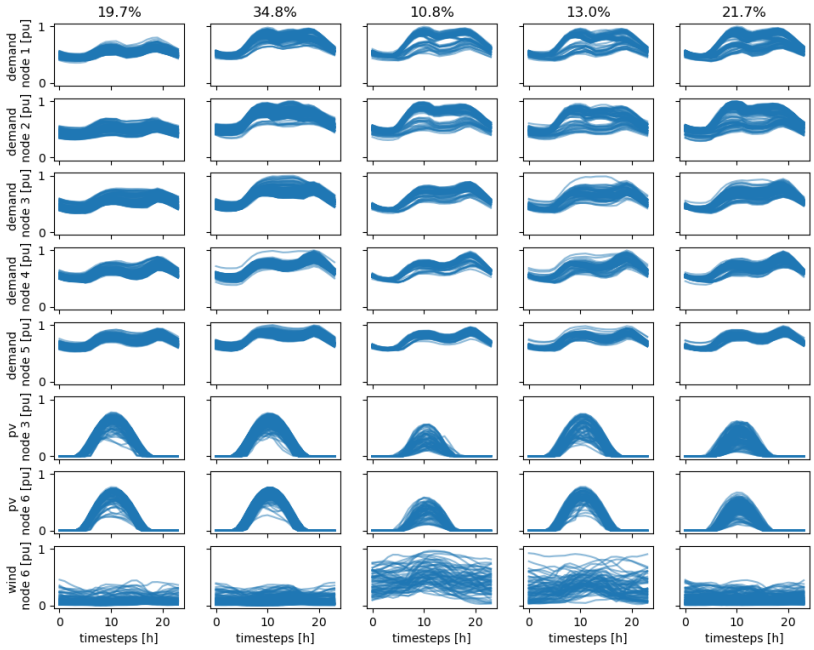

Figure 6-13: Result of clustering the 40 climate-year scenario's into daily profiles, when reducing each daily scenarios to its twelve characteristics. Each column shows the demand, pv and wind profiles per cluster. Above each column the probability of each cluster is indicated.

Different clusters are however obtained when applying the feature reduction method outlined in section 5.2.2, i.e. by first clustering the node dimension. The result is shown in Figure 6-14.

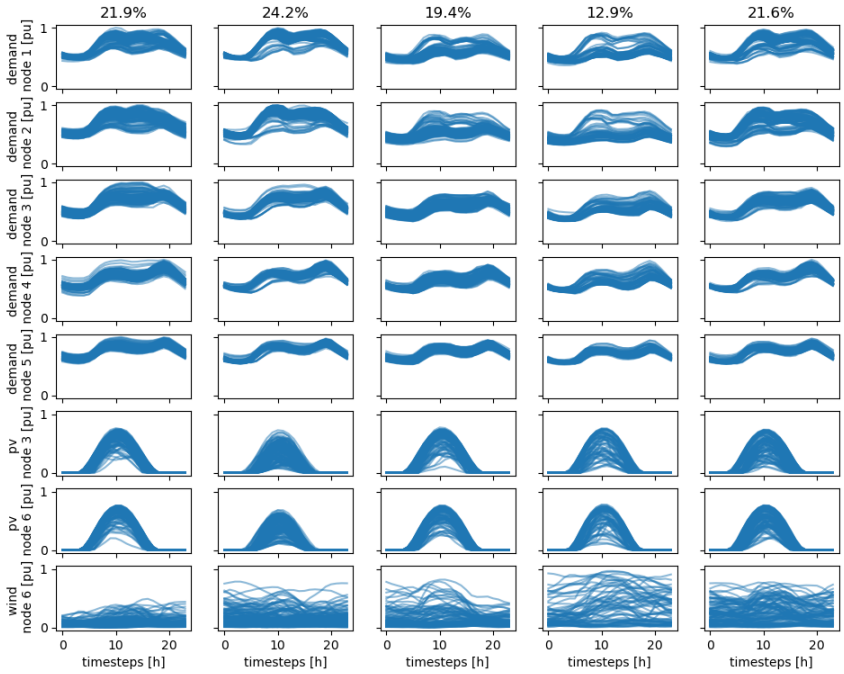

Figure 6-14: Result of clustering the 40 climate-year scenario's into daily profiles, when reducing after having first reduced the features by clustering on the node dimension. Each column shows the demand, pv and wind profiles per cluster. Above each column the probability of each cluster is indicated.

In this case, it is not obvious from the clustering results itself which clusters are most significant for the planning tool optimisation problem: the clusters as shown in Figure 6-13 or the clusters as shown in Figure 6-14.

Therefore, within the planning tool testing activities, the planning tool decisions will be compared when fed with different scenario inputs, obtained using different feature reduction methods, to determine which feature reduction method performs best.

7. Conclusion

The scenario generation and reduction approach developed in this task uses pan-EU macro-scenarios as an input to generate operational scenarios for the stochastic inputs, namely vRES and hydropower generation, electricity demand and power plant availabilities by applying a Monte Carlo approach. Subsequently, the generated operational scenarios / MC years are reduced by a clustering approach. The developed methodology consists of the following steps:

- A time series generator, sampling a broad variety of possible realizations of vRES and hydropower generation based on historical meteorological and hydrological data
- A time series generator, which samples temperature-sensitive load profiles based on sensitivity factors depending on outdoor ambient temperature
- A heuristic, scheduling planned maintenance on a yearly basis
- A heuristic, generating randomized thermal power plant outages and hourly availability profiles
- A clustering algorithm, identifying and selecting a reduced sub-set of relevant operational scenarios

An extensive literature review has been carried out to identify MC scenario generation and reduction techniques that are compatible with FlexPlan's time-series based planning approach. Current literature and especially adequate data sources have been reviewed on each of the points mentioned above.

As a consequence, significant work has been done on the identification of suitable datasets for modelling the uncertain inputs, namely vRES and hydropower generation and demand. Furthermore, spatial and temporal correlations in vRES generation time series of 40 historical years have been analyzed. Regarding wind and solar generation, a database containing hourly capacity factors for roughly 290 locations (NUTS-2-regions) in Europe for in total 40 historic years has been created (cf. section 4.2.1). Additionally, uncertain phenomena in historical hydropower generation and demand time series of 36 historic years have been analyzed in depth. With regard to hydropower generation and load, two additional databases have been created containing hourly hydropower and load capacity factors for roughly 50 locations (market areas) in Europe, for 36 historic years in total (cf. section 4.2.2 and section 4.2.3). Considering the identified spatial and temporal correlations of vRES feed-in, hydropower generation and demand, a scenario generation approach based on detailed hourly capacity factors has been developed (cf. section 4.3).

Considering the number of 40 operational scenarios per macro-scenario and target year, significant work has been done on the conceptual design of an adequate scenario reduction technique. The developed scenario reduction approach makes use of a clustering algorithm.

Finally, the developed scenario generation and reduction method has been applied to a small-scale numerical example (6-node test case) for testing purposes. For this test case, 40 scenario years have been generated, and the set of 40 years has been reduced to (1) 5 yearly scenarios and (2) 5 daily scenarios. The scenario reduction exercise shows that the years within the original generated set do not differ significantly, therefore the clusters of yearly profiles identified within the generated set are also very similar. In contrast to this, the differences between the daily scenarios are much larger, and the produced clusters also differ notably.

The different proposed feature reduction methodologies have been applied to the generated scenario set. Especially the yearly scenarios, have (too) many features when applying clustering without feature reduction.

It is however not straightforward to decide which feature reduction method produces the best results, i.e. the planning tool produces optimal decisions. Therefore, within the planning tool testing activities, the planning tool output will be compared when fed with different options of scenario clustering method as well as cluster sizes. This comparison will decide which feature reduction method to apply for the large regional case planning optimizations targeted within the FlexPlan project.

8. References

[1] C. Spieker, "Europäische Strommarkt- und Übertragungsnetzsimulation zur techno-ökonomischen Bewertung der Netzentwicklung", Doctoral dissertation, Technische Universität Dortmund, 2019.

[2] V. Liebenau, "Einfluss der Regionalisierung Erneuerbarer Energien sowie innovativer Konzepte auf die Netzentwicklungsplanung", Doctoral dissertation, Technische Universität Dortmund, 2018.

[3] J. Teuwsen, "Gegenüberstellung divergenter Zukunftsszenarien des Energieversorgungssystems", Doctoral dissertation, Technische Universität Dortmund, 2016.

[4] FLEXPLAN Consortium, "D4.1 Pan-European scenario data", https://flexplan-project.eu/wp-content/uploads/2020/08/D4.1_20200803_V1.pdf

[5] ENTSO-E TYNDP 2020 Scenarios

[6] Deutscher Wetterdienst, „Numerical weather prediction models," Online Available: https://www.dwd.de/EN/research/weatherforecasting/num_modelling/01_num_weather_prediction_modells/num_weather_prediction_models_node.html

[7] S. Lumbreras and A. Ramos, "*The new challenges to transmission expansion planning. Survey of recent practice and literature review,*" Electric Power Systems Research, vol. 134, pp. 19–29, May 2016.

[8] M.V.F. Pereira, L.M.V.G. Pinto, S.H.F. Cunha, G.C. Oliveira, *A decomposition approach to automated generation/transmission expansion planning*, IEEE Trans. Power Apparatus Syst. PAS-104 (1985) 3074–3083.

[9] J.F. Benders,"*Partitioning Procedures for Solving Mixed Variables Programming Problems,*" Numerische Mathematik 4, pp. 238-252, 1962.

[10] O. Tor, A. Guven, M. Shahidehpour, "*Congestion-driven transmission planning considering the impact of generator expansion*", IEEE Transactions on Power Systems, 2008.

[11] F. Cadini, E. Zio, C. Petrecu, "*Optimal expansion of an existing electrical power transmission network by multi-objective genetic algorithms*", Reliability Engineering and System Safety, 2010.

[12] FA. Sousa, E. Asada, "*Long-term transmission system expansion planning with multi-objective evolutionary algorithm*", Electric Power Systems Research, 2015.

[13] G. Latorre, R. Cruz, I. Areiza, A. Villegas, "*Classification of publications and models on transmission expansion planning*", IEEE Transactions on Power Systems, 2003.

[14] ENTSO-E, *Mid-term Adequacy Forecast 2019*, https://www.entsoe.eu/outlooks/midterm/

[15] German TSO, "*Szenariorahmen für den Netzentwicklungsplan Strom 2030 (Version 2019) – ÜNB-Entwurf*", January 2018, https://www.netzentwicklungsplan.de/sites/default/files/paragraphs-files/%C3%9CNB-Entwurf_Szenariorahmen_2030_V2019_0_0.pdf

[16] Ansari, Dawud and Holz, Franziska and al-Kuhlani, Hashem, *Energy Outlooks Compared: Global and Regional Insights* (December 12, 2019). DIW Berlin Discussion Paper No. 1837 (2019), Available at SSRN: https://ssrn.com/abstract=3501509 or http://dx.doi.org/10.2139/ssrn.3501509

[17] C.-L. Su, "*Probabilistic load-flow computation using point estimate method*," in Proc. IEEE Trans. Power Syst., vol. 20, no. 4, S. 1843-1851, Nov. 2005.

[18] J. M. Morales und J. Pérez-Ruiz, "*Point estimate schemes to solve the probabilistic power flow,*" in Proc. IEEE Trans. Power Syst., vol. 22, no. 4, S. 1594-1601, Nov. 2007.

[19] M. Mohammadi, "*Probabilistic harmonic load flow using fast point estimate method,*" in IET Gener. Transm. Distrib., vol. 9, nr. 13, S. 1790-1799, Jan. 2015.

[20] A. Leite da Silva, W. S. Sales, L.A. da Fonseca Manso, R. Billinton, „*Long-Term Probabilistic Evaluation of Operating Reserve Requirements With Renewable Sources*", In: IEEE Trans. Power Syst., vol. 25, no. 1, S. 106–116, 2010.

[21] D. Villanueva, A. Feijóo, J. L. Pazos "*Probabilistic Load Flow Considering Correlation between Generation, Loads and Wind Power*", Smart Grid and Renewable Energy, vol. 2, no. 1, S.12-20 Februar, 2011.

[22] Edgar Nuño Martinez, Nicolaos Cutululis, Poul Sørensen, "*High dimensional dependence in power systems: A review,*" Renewable and Sustainable Energy Reviews, Volume 94, 2018, Pages 197-213.

[23] R. B. Nelsen, "*An Introduction to Copulas*", 2. Edition, Springer Verlag, 2006.

[24] T. Shi, L. Zhu, X. Wang, "*Research on correlation evaluation and joint probability model of multi-wind generation scenarios*"; IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies, Changsha, China, 2015.

[25] H. Louie; "*Evaluating Archimedean Copula Models of Wind Speed for Wind Power Modeling,*" IEEE Power & Energy Society Conference and Exposition in Africa, Johannesburg, South Africa, 2012.

[26] Y. Li, W. Li, K. Xie, "*Modelling wind speed dependence in system reliability assessment using copulas,*" in IET Renewable Power Generation, vol. 6, S. 392–399, 2012.

[27] H. Park, R. Baldick, D. Morton, "*A Stochastic Transmission Planning Model With Dependent Load and Wind Forecasts,*" in IEEE Trans. Power Syst., vol. 30, no. 6, S. 3003–3011, 2015.

[28] J. H. Roh, M. Shahidehpour, L. Wu, "*Market-Based Generation and Transmission Planning With Uncertainties*",in IEEE Trans. Power Syst., vol. 24, no. 3, 2009.

[29] C. Wan, Z. Y. Dong, F. Luo, M. Lu, P. Zhang, K. P. Wong, „*Cumulant-Based Probabilistic Load Flow Calculation in a Market Environment*", International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), Shandong, China, 2011.

[30] A. Schellenberg, W. Rosehart, J. Aguado, "*Cumulant-Based Probabilistic Optimal Power Flow (P-OPF) With Gaussian and Gamma Distributions*", in IEEE Trans. Power Syst., vol. 20, no. 2, 2005

[31] L. Xue, L. Yuzeng, Z. Shaohua, "*Analysis of Probabilistic Optimal Power Flow Taking Account of the Variation of Load Power*",In: IEEE Trans. Power Syst., vol. 23, no. 3, S. 992-999, 2008.

[32] C.A. Correa Florez, R.A. Bolaños Ocampo, A.H. Escobar Zuluaga, "*Multi-objective transmission expansion planning considering multiple generation scenarios*", In: International Journal of Electrical Power & Energy Systems 62, S. 398–409, 2014.

[33] M. C. da Rocha and J. T. Saraiva, "*Transmission expansion planning — A multiyear PSO based approach considering load uncertainties,*" 2013 IEEE Grenoble Conference, Grenoble, 2013, pp. 1-6, doi: 10.1109/PTC.2013.6652080.

[34] D. Heinemann, J. Parisi, H.-P. Waldl und H. G. Beyer: *Energiemeteorologie*. Bd. Physikalische Blätter 55. Nr. 4. Weinheim: WILEY-VCH Verlag GmbH, 1999. doi: 10.1002/phbl.19990550413.

[35] E. Hau: *Windkraftanlagen, Grundlagen, Technik, Einsatz, Wirtschaftlichkeit*. Springer-Verlag GmbH, 11. Okt. 2014. isbn: 3642288766. doi: 10.1007/978-3-642-28877-7.

[36] G. Gross, T. Frey und P. Trute: *Die Anwendung numerischer Simulationsmodelle zur Berechnung der lokalen Windverhältnisse in komplexem Gelände*. Bd. Nr. 20. DEWI-Magazin. Wilhelmshaven, Feb. 2002.

[37] G. Groß und C. Etling: *Numerische Simulationsmodelle*. Bd. Jahrg. 30, Nr. 1/2. promet. Deutscher Wetterdienst 2003, Nov. 2003.

[38] S. Pfenninger und I. Staffell: *Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data*. Bd. 114. Elsevier BV, Nov. 2016, S. 1251–1265. doi: 10.1016/j.energy.2016.08.060.

[39] I. Staffell und S. Pfenninger: *Using bias-corrected reanalysis to simulate current and future wind power output*. Bd. 114. Elsevier BV, Nov. 2016, S. 1224–1239. doi: 10.1016/j.energy.2016.08.068.

[40] L. Lück, M. Ketov, J. *Unland und A. Moser: Modellierung der Einspeisung von Windenergie- und Photovoltaikanlagen für Strommarkt- und Netzbetriebssimulationen*. 10. Internationale Energiewirtschaftstagung IEWT 2017. 2017.

[41] G. Schubert: *Modellierung der stündlichen Photovoltaik- und Windstromeinspeisung in Europa*. 12. Symposium Energieinnovation - TU Graz. 2012.

[42] A. Wagner: *Photovoltaik Engineering, Handbuch für Planung, Entwicklung und Anwendung*. 4. Auflage. Springer Heidelberg Dordrecht London New York, 2015. isbn: 978-3-662-48640-5. doi: 10.1007/978-3-662-48640-5.

[43] Jourdier, B. "*Evaluation of ERA5, MERRA-2, COSMO-REA6, NEWA and AROME to simulate wind power production over France,*" Advances in Science and Research, Volume 17, 2020, p. 63 – 77, Online available: https://doi.org/10.5194/asr-17-63-2020

[44] Kaspar, F. et al. "*Regional atmospheric reanalysis activities at Deutscher Wetterdienst: review of evaluation results and application examples with a focus on renewable energy,*" Advances in Science and Research, Volume 17, 2020, p. 115 – 128,Online available: https://doi.org/10.5194/asr-17-115-2020

[45] Kiss P, Varga L, Janosi IM. *Comparison of wind power estimates from the ECMWF reanalyses with direct turbine measurements*. J Renew Sustain Energy 2009;1(3):033105e11.

[46] Kubik M, Brayshaw DJ, Coker PJ, Barlow JF. *Exploring the role of reanalysis data in simulating regional wind generation variability over Northern Ireland*. Renew Energy 2013;57:558e61.

[47] Staffell I, Green R. *How does wind farm performance decline with age?* Renew Energy 2014;66:775e86. URL, http://www.sciencedirect.com/science/article/pii/S0960148113005727.

[48] Cannon DJ, Brayshaw DJ, Methven J, Coker PJ, Lenaghan D. *Using reanalysis data to quantify extreme wind power generation statistics: a 33 year case study in Great Britain*. Renew Energy 2015;75(0):767e78.

[49] Drew DR, Cannon DJ, Brayshaw DJ, Barlow JF, Coker PJ. *The impact of future offshore wind farms on wind power generation in Great Britain*. Resources 2015;4(1):155e71. URL, http://www.mdpi.com/2079-9276/4/1/155.

[50] Sharp E, Dodds P, Barrett M, Spataru C. *Evaluating the accuracy of CFSR reanalysis hourly wind speed forecasts for the UK, using in situ measurements and geographical information*. Renew Energy 2015;77:527e38. URL, http://www.sciencedirect.com/science/article/pii/S0960148114008520.

[51] Andresen GB, Søndergaard AA, Greiner M. *Validation of danish wind time series from a new global renewable energy atlas for energy system analysis*. 2014. arXiv:1409.3353 [physics]arXiv: 1409.3353. URL, http://arxiv.org/abs/1409.3353.

[52] Olauson J, Bergkvist M. *Modelling the Swedish wind power production using MERRA reanalysis data*. Renew Energy 2015;76(0):717e25.

[53] Lu X, McElroy MB, Kiviluoma J. *Global potential for wind-generated electricity*. Proc Natl Acad Sci 2009;106(27):10933e8.

[54] Cosseron A, Gunturu UB, Schlosser CA. *Characterization of the wind power resource in Europe and its intermittency*. Energy Procedia 2013;40(0):58e66.

[55] Huber M, Dimkova D, Hamacher T. *Integration of wind and solar power in Europe: assessment of flexibility requirements*. Energy 2014;69:236e46. Online available at: http://linkinghub.elsevier.com/retrieve/pii/S0360544214002680.

[56] Becker S, Frew BA, Andresen GB, Zeyer T, Schramm S, Greiner M, et al. "*Features of a fully renewable US electricity system: optimized mixes of wind and solar PV and transmission grid extensions*". Energy 2014;72(0):443e58.

[57] McKenna R, Hollnaicher S, v. d. Leye P Ostman, Fichtner W. *Cost-potentials for large onshore wind turbines in Europe.* Energy 2015;83(0):217e29.

[58] Gunturu UB, Schlosser CA. *Characterization of wind power resource in the United States*. Atmos Chem Phys 2012;12(20):9687e702.

[59] Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, et al. MERRA: *NASA's modern-era retrospective analysis for research and applications.* J Clim 2011;24(14):3624e48. URL, http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00015.1

[60] Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, et al. *The ERA-Interim reanalysis: configuration and performance of the data assimilation system.* Q J R Meteorol Soc 2011;137(656):553e97. URL, http://onlinelibrary.wiley.com/doi/10.1002/qj.828/abstract.

[61] Kobayashi S, Ota Y, Harada Y, Ebita A, Moriya M, Onoda H, et al. *The JRA-55 reanalysis: general specifications and basic characteristics.* J Meteorol Soc Jpn Ser II 2015;93(1):5e48.

[62] IRENA. *Renewable energy capacity statistics 2015*. 2015. URL, http://www.irena.org/DocumentDownloads/Publications/IRENA_RE_Capacity_Statistics_2015.pdf.

[63] Haller M, Ludig S, Bauer N. *Decarbonization scenarios for the EU and MENA power system: considering spatial distribution and short term dynamics of renewable generation*. Energy Policy 2012;47:282e90. URL, http://www.sciencedirect.com/science/article/pii/S0301421512003746.

[64] Heide D, von Bremen L, Greiner M, Hoffmann C, Speckmann M, Bofinger S. *Seasonal optimal mix of wind and solar power in a future, highly renewable Europe*. Renew Energy 2010;35(11):2483e9. URL, http://www.sciencedirect.com/science/article/pii/S0960148110001291.

[65] Jurus P, Eben K, Resler J, Krc P, Kasanický I, Pelikan E, et al. *Estimating climatological variability of solar energy production.* Sol Energy 2013;98: 255e64. URL, http://www.sciencedirect.com/science/article/pii/S0038092X13004088.

[66] Janker KA. *Aufbau und Bewertung einer für die Energiemodellierung verwendbaren Datenbasis an Zeitreihen erneuerbarer Erzeugung und sonstiger Daten*. PhD thesis. München: Technische Universität München; 2015. URL, https://mediatum.ub.tum.de/doc/1207265/1207265.pdf

[67] e-HIGHWAY 2050 D8.2.a *Enhanced methodology for the computation of Demand/Generation scenarios*

[68] Troccoli, A., Goodess, C., Jones, P., Penny, L., Dorling, S., Harpham, C., Dubus, L., Parey, S., Claudel, S., Khong, D.-H., Bett, P. E., Thornton, H., Ranchin, T., Wald, L., Saint-Drenan, Y.-M., De Felice, M., Brayshaw, D., Suckling, E., Percy, B., and Blower, J.: *Creating a proof-of-concept climate service to assess future renewable energy mixes in Europe: An overview of the C3S ECEM project*, Adv. Sci. Res., 15, 191–205, https://doi.org/10.5194/asr-15-191-2018, 2018.

[69] J. P. Bukenberger , M. D. Webster, *"Approximate Latent Factor Algorithm for Scenario Selection and Weighting in Transmission Expansion Planning"* IEEE Trans. on Power Systems, Vol.. 35, No. 2, March 2020, pp 1099-1108.

[70] R. Loulou, U. Remne, A. Kanudia, A. Lehtila, and G. Goldstein, "*Documentation for the TIMES model: Part I*," ETSAP, Apr. 2005.4

[71] P. Nahmmacher, E. Schmid, and B. Knopf, "*Documentation of LIMES-EU—A long-term electricity system model for Europe*," Potsdam Inst. Climate Impact Res., Potsdam, Germany, Tech. Rep., 2014.

[72] GARPUR Consortium, "D4.2, Upgrading of the decision-making process for system development". Garpur, FP7-ENERGY-2013-1, 2016.

[73] K. Poncelet, H. Höschle, E. Delarue, A. Virag, W. D'haeseleer," *Selecting Representative Days for Capturing the Implications of Integrating Intermittent Renewables in Generation Expansion Planning Problems*". IEEE Trans. on Power Systems, Vol.. 32, No. 3, May 2017, pp. 1936-1947

[74] R. Alvarez, A. Moser, C. A. Rahmann, "*Novel Methodology for Selecting Representative Operating Points for the TNEP*", IEEE Trans. on Power Systems, Vol.. 32, No. 3, May 2017, pp. 2234-2242

[75] P Nahmmacher, E. Schmid, L. Hirth, B. Knopf, "*Carpe diem: A novel approach to select representative days for long-term power system models with high shares of renewable energy sources*", 2014. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2537072

[76] H. Yu, C. Y. Chung, K. P. Wong, H. W. Lee, J. H. Zhang, "*Probabilistic Load Flow Evaluation With Hybrid Latin Hypercube Sampling and Cholesky Decomposition*", IEEE Trans. on Power Systems, Vol.. 24, No. 2, May 2009, pp. 661-667.

[77] W. Bukhsh, K. Bell, A. Vergnol, A. Weynants and J. Sprooten, "*Enhanced, Risk-Based System Development Process: A Case Study from the Belgian Transmission Network*," 2018 Power Systems Computation Conference (PSCC), Dublin, 2018, pp. 1-9.

[78] Steinley, D., & Brusco, M. J. (2007). *Initializing k-means batch clustering: A critical evaluation of several techniques.* Journal of Classification, 24(1), 99–121.

[79] J. Quintero, H. Zhang, Y. Chakhchoukh, V. Vittal, and G. T. Heydt, "*Next generation transmission expansion planning framework: Models, tools, and educational opportunities*," IEEE Trans. Power Syst., vol. 29, no. 4, pp. 1911–1918, Jul. 2014.

[80] E. Shayesteh, B. F. Hobbs, and M. Amelin, *"Scenario reduction, network aggregation, and DC linearisation: Which simplifications matter most in operations and planning optimisation*?" IET Gener. Transmiss. Distrib., vol. 10, no. 11, pp. 2748–2755, 2016.

[81] A. Papavasiliou and S. S. Oren, "*Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network*," Oper. Res., vol. 61, no. 3, pp. 578–592, May 2013.

[82] K. Poncelet, E. Delarue, D. Six, J. Duerinck, andW. D'haeseleer, "*Impact of the level of temporal and operational detail in energy-system planning models*," Appl. Energy, vol. 162, pp. 631–643, Jan. 2016

[83] C. Yuan, C. Gu, F. Li, B. Kuri, and R. Dunn, "*New problem formulation of emission constrained generation mix*," IEEE Trans. Power Syst., vol. 28, no. 4, pp. 4064–4071, Nov. 2013.

[84] A. Belderbos and E. Delarue, "*Accounting for flexibility in power system planning with renewables*," Int. J. Elect. Power Energy Syst., vol. 71, pp. 33–41, Oct. 2015.

[85] W. Omran, M. Kazerani, and M. Salama, "*A clustering-based method for quantifying the effects of large on-grid PV systems*," IEEE Trans. Power Del., vol. 25, no. 4, pp. 2617–2625, Oct. 2010.

[86] M. Nick, R. Cherkaoui, and M. Paolone, "*Optimal allocation of dispersed energy storage systems in active distribution networks for energy balance and grid support*," IEEE Trans. Power Syst., vol. 29, no. 5, pp. 2300–2310, Sep. 2014.

[87] K. Bruninx and E. Delarue, "*Scenario reduction techniques and solution stability for stochastic unit commitment problems*", Proc. IEEE Int. Energy Conf., pp. 1-7, Apr. 2016

[88] J. Dupacova, N. Gröwe-Kuska, and W. Römisch, "*Scenario reduction in stochastic programming*," Mathematical Programming, vol. 95, no. 3, pp. 493–511, 2003.

[89] J. M. Morales, S. Pineda, A. J. Conejo, and M. Carrion, "*Scenario reduction for futures market trading in electricity markets*," IEEE Trans. Power Syst., vol. 24, no. 2, pp. 878–888, 2009

[90] Aggarwal CC, Hinneburg A, Keim DA (2001) *On the surprising behavior of distance metrics in high dimensional spaces*. In: ICDT, London, pp 420–434

[91] A. K. Jain, M. N. Murty, and P. J. Flynn. "*Data clustering: a review*." ACM Comput. Surv. 31, 3 (Sept. 1999), 264–323.

[92] A. Fahad *et al.*, "*A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis*," in IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, Sept. 2014.

[93] https://scikit-learn.org/stable/index.html

[94] Jolliffe Ian T. and Cadima Jorge, "*Principal component analysis: a review and recent developments*" Phil. Trans. R. Soc. A., Vol. 374, Issue 2065.

[95] 2016.ENTSO-E *TYNDP 2018 Transmisson Grid Dataset*, https://docstore.entsoe.eu/stum/

[96] Pfenninger, Stefan and Staffell, Iain: www.renewables.ninja

[97] Lauret P, Boland J, Ridley B. *Bayesian statistical analysis applied to solar radiation modelling*. Renew Energy 2013;49:124e7. URL, http://www.sciencedirect.com/science/article/pii/S0960148112000602

[98] Ridley B, Boland J, Lauret P. *Modelling of diffuse solar fraction with multiple predictors*. Renew Energy 2010; 35(2):478e83. URL, http://www.sciencedirect.com/science/article/pii/S0960148109003012

[99] IEC 61724:1998 IEC. *Photovoltaic system performance monitoring – guidelines for measurement, data exchange and analysis*. 1998. URL, https://law.resource.org/pub/in/bis/S05/is.iec.61724.1998.pdf

[100] I. Staffell and S. Pfenninger, 2016. *Using Bias-Corrected Reanalysis to Simulate Current and Future Wind Power Output. Energy*, 114, 1224–1239. http://dx.doi.org/10.1016/j.energy.2016.08.068

[101] ENTSO-E *Mid-term Adequacy Forecast 2019 Edition - Hydropower modelling – New database complementing PECD V.1.0,* Online available: https://eepublicdownloads.entsoe.eu/clean-documents/sdc-documents/MAF/2019/Hydropower_Modelling_New_database_and_methodology.pdf

[102] ENTSO-E *Hydropower modelling data (PECD) in CSV format,* Online available: https://doi.org/10.5281/zenodo.3929772

[103] European Hydrological Predictions for the Environment and consists of high resolution (E-HYPE), https://hypeweb.smhi.se/explore-water/geographical-domains/#europehype

[104] Clim4Energy – Copernicus Project, https://clim4energy.climate.copernicus.eu/

[105] ENSTO-E *Mid-term Adequacy Forecast 2018 – Appendix 1: Methodology and Detailed Results, 2018 edition,* Online available: https://eepublicdownloads.entsoe.eu/clean-documents/sdc-documents/MAF/2018/MAF%202018%20Methodology%20and%20Detailed%20Results.pdf

[106] ENTSO-E *Winter Outlook 2019/2020 and Summer Review 2019*, Online available: https://www.entsoe.eu/outlooks/seasonal/ Data set online available: https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/sdc-documents/seasonal/SOR2020/data/Demand_SOR%2020.xlsx

[107] M. Doquet; C. Fourment; J. Roudergues, "*Generation & transmission adequacy of large interconnected power systems: A contribution to the renewal of Monte-Carlo approaches,*" PowerTech, 2011 IEEE Trondheim, June 19-23, 2011

[108] L. Garver, "*Transmission network estimation using linear programming,*" Power Apparatus and Systems, IEEE Transactions on, vol. PAS-89, no. 7, pp. 1688 –1697, sept. 1970. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4074249

[109] K. Yata, M. Aoshima, "*Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix*", Journal of Multivariate Analysis, Volume 101, Issue 9, 2010, pp 2060-2077.