Frysztacki *et al. Energy Informatics* 2022, **5**(1):4
https://doi.org/10.1186/s42162-022-00187-7

Energy Informatics

METHODOLOGY Open Access

A comparison of clustering methods for the spatial reduction of renewable electricity optimisation models of Europe

Martha Maria Frysztacki[1*], Gereon Recht[1] and Tom Brown[1,2]

*Correspondence:
martha.frysztacki@kit.edu

[1] Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
[2] Institute of Energy Technology, Technical University of Berlin, Berlin, Germany

Abstract

Modeling the optimal design of the future European energy system involves large data volumes and many mathematical constraints, typically resulting in a significant computational burden. As a result, modelers often apply reductions to their model that can have a significant effect on the accuracy of their results. This study investigates methods for spatially clustering electricity system models at transmission level to overcome the computational constraints. Spatial reduction has a strong effect both on flows in the electricity transmission network and on the way wind and solar generators are aggregated. Clustering methods applied in the literature are typically oriented either towards preserving network flows or towards preserving the properties of renewables, but both are important for future energy systems. In this work we adapt clustering algorithms to accurately represent both networks and renewables. To this end we focus on hierarchical clustering, since it preserves the topology of the transmission system. We test improvements to the similarity metrics used in the clustering by evaluating the resulting regions with measures on renewable feed-in and electrical distance between nodes. Then, the models are optimised under a brownfield capacity expansion for the European electricity system for varying spatial resolutions and renewable penetration. Results are compared to each other and to existing clustering approaches in the literature and evaluated on the preciseness of siting renewable capacity and the estimation of power flows. We find that any of the considered methods perform better than the commonly used approach of clustering by country boundaries and that any of the hierarchical methods yield better estimates than the established method of clustering with *k*-means on the coordinates of the network with respect to the studied parameters.

Keywords: Energy system modelling, Open source modelling, Spatial network clustering, Hierarchical clustering, Electrical distance, K-means clustering, Clustering by political borders, Renewable energy, Power flow

Introduction

Electricity system planners typically use optimisation models to design the combinations of generation, storage and transmission that meet different climate objectives, such as CO_2 reduction scenarios, limiting the temperature increase or phasing out nuclear

energy. They are motivated by political goals, such as the Paris Agreement (2012) or the European Green Deal (2019). Many of these goals require a high share of renewable generation.

An energy system model suited for such modelling tasks must capture spatio-temporal variations of both renewable resources and electricity demand as well as network bottlenecks, which are already constraining today. Hence, the models must have a high spatial and temporal resolution. One approach for achieving this is to embed historical data into the model that contains hourly observations for sites at every few dozen kilometers, such as provided by the European Centre for Medium-Range Weather Forecasts (ECMWF): ERA5 Reanalysis (2020) or Pfeifroth et al. (2017). However, a sufficient resolution entails large amounts of data, which leads to significant computational burdens. They arise because the modeling task is typically formulated as an optimisation problem which is subject to many mathematical constraints. These constraints account for the physics of the system, such as an accurate representation of power flows or (renewable) generation. To overcome the computational burdens, different approaches are established in the literature. They can be manifold ranging from linearisation to multi-level approaches combining aggregation and decomposition methods. An overview on potentials to reduce model complexity is provided in Kotzur (2021). However, applying a linearisation to a large-scale mathematical model of the European electricity system is not sufficient to obtain a computational feasibility. The remaining option is to reduce the model size in its temporal or spatial dimension using aggregation methods.

Temporal clustering methods and their impact on the optimal energy system design are already well analysed (Kotzur et al. 2018). The main findings of previous studies include the need for at least hourly modelling resolution (Pfenninger et al. 2014) and the need to include extreme weather events (Perera et al. 2020). On the spatial side, many studies pursue the approach of either using the full electricity substation level resolution for the transmission grid but only in selected regions (Sasse and Trutnevyte 2020), reducing the full model to a smaller equivalent using clustering methods (eHighways 2050 Final Reports 2015; Neumann 2021; Zeyen et al. 2020; Tröndle et al. 2020) or both (Frysztacki and Brown 2020; Lombardi et al. 2020) and make suggestions for the future energy system or the modelling process based on the results obtained by the reduced models. However, to take advantage of trade over large distances, the model should cover the total area of political interest, which typically includes at least a whole continent, not just single regions (Tröndle et al. 2020; Schlachtberger et al. 2017). In case of clustering the model, it is ongoing research to find which method suits which research application and how precise reduced model results are compared to solutions obtained from higher resolved models.

This study aims to address the issue of spatial exactness of different clustering methods by extending recently proposed solutions and comparing their performance in the application of electricity system models. We extend previous methods to account for the spatio-temporal availability of renewable resources and incorporate considerations of the network topology and electrical connectivity. Results obtained from reduced models based on the different aggregation algorithms are compared against each other and against established reduction methods from the literature. The obtained low-resolution estimates are evaluated against an accurate power flow and the siting of renewable

capacities and associated storage options (solar, wind, battery and hydrogen) obtained from higher-resolution simulations.

To increase transparency of the results, we use the open model PyPSA-Eur (Brown et al. 2018; Hörsch et al. 2018) which builds on an open database.

State of the art

In the recent energy system modeling (ESM) literature, suggested solutions to spatially reduce high-resolution models to smaller equivalents include different techniques that focus on individual features of the system. These solutions can be categorised by whether they focus on (i) representing the network or (ii) the variability of renewable resources.

(i) Approaches that focus on the network representation and therefore on accurately approximating power flows include the following methods: the Ward equivalent (Ward 1949), a hybrid method consisting of *k*-means and an evolutionary algorithm (Cotilla-Sanchez et al. 2013), clustering into zones based on the similarity of the power transfer distribution factors (PTDFs) (Shi and Tylavsky 2015), *k*-medoids operating on a combination of electrical parameters of the grid as well as their geographical length (2015), spectral partitioning taking into account the available transfer capability (ATC) (Shayesteh et al. 2017) or density-based hierarchical clustering operating on the lines reactance (Biener and Garcia Rosas 2020). All these approaches use distance or similarity measures on electrical parameters of the transmission grid, often referred to as electrical distance. These methods are designed for a good approximation of power flows and mostly evaluated comparing the power flows of a highly resolved model to the one of a reduced model without changing the generation fleet. However, power flows are strongly impacted when moving away from conventional generation to other resources as shown in Shi et al. (2012) (a study conducted with the Ward equivalent) for the example of switching from coal fired electricity generation to natural gas. Therefore it remains unclear if these methods are applicable when moving towards high shares of renewables as they are not designed to precisely approximate the spatio-temporal variability of wind and solar. This is especially true for models where the final installed capacity as well as its spatial distribution is subject to optimisation, such that no a-priori estimate of power flows can be made.

(ii) On the other hand, techniques that focus on an accurate representation of renewables include hierarchical clustering applied on a database of electricity demand, conventional generation and renewable profiles (Kueppers et al. 2020), max-p-regions applied on a database of wind and solar potential (Fleischer 2020) or a combination of *k*-means++ with the max-p-regions algorithm applied separately on the full load hours of wind, solar and electricity demand (Siala and Mahfouz 2019). Radu et al. (2021) proposes a novel screening routine that identifies relevant generation sites to be passed to a capacity expansion problem. All these methods include either a synthesised transmission grid or one at very low resolution. The downside of such approaches is that transmission bottlenecks within large regions can not be identified, and therefore power flows within large regions are not considered at all. By ignoring the grid, transmission congestion hinders exploitation of the best available resource sites and results of these models may not be feasible in reality.

A clustering method that is neutral to both of these features is to reduce the model by applying *k*-means on the coordinates of the network nodes (Frysztacki et al. 2021). However, location-wise clustering has no inevitable correlation with either the transmission grid nor the renewable resources, hence requiring relatively large spatial resolutions to yield good results. Furthermore, using *k*-means on locations ignores the connectivity of the grid, and could end up aggregating two nodes that were previously disconnected, resulting in strong distortion of the network representation.

Research gap

In the present contribution we focus on two relevant settings: Improvements in the clustering process that can capture both the important features to accurately portray renewable generation while incorporating the transmission network and evaluating the proposed methods on both the representation of renewable generation and power flows. We define metrics to determine if good renewable generation sites after the clustering are maintained while incorporating the electricity grid by aggregating only nodes connected by an existing transmission line using Ward's method (hierarchical clustering). This approach is completely novel in the context of ESM. For this method, we distinguish between two features: The aggregated quantity of annual capacity factors (similar to Siala and Mahfouz (2019) with the adaptation of incorporating the grid), and the full time series of renewables (a novel metric employed in the context of ESM). Using the full time series to define regions is motivated by the fact that regions with similar capacity factors may have very different time profiles, depending on how their generation is correlated over space; by using the full time series, we avoid aggregating sites with very different profiles.

We compare all results obtained by the same model using the same experimental setup and input data. This increases transparency and guarantees that differences in the results occur because of the clustering process, not because of differences in the data or other parameters of the model. To complete the comparison, we include two common reduction methods from the literature: *k*-means clustering on the coordinates of the network and clustering based on country borders.

Outline of the paper

The remainder of the paper is as follows: First, we introduce notation, the applied data sets and the set-up of the model (chapter Notation, Data and Model Set-Up). Second, we introduce a pre-aggregation method on a subset of network nodes that reduces the initial network size by a factor of approximately two. Then we present the application of the following clustering methods to energy system models for further model reduction: *k*-means, a benchmark clustering technique based on the coordinates of the network nodes that was used in several publications in the past; Ward's method, for which we adapt the metric to a time-aggregated annual and on a time-resolved hourly feature of the system; and Modularity Maximization, that involves considerations of electrical parameters of the model. (Sub-chapter in Clustering Methodology). At lowest model resolution, the aggregated networks are equivalent to a second benchmark aggregation method that represents each country and synchronous zone with one node. All methods (except for the benchmark methods) are of

hierarchical nature, because this approach takes into account the network topology. Nonetheless, different similarity (capacity factors and time-series) or distance (electrical distance) measures are applied for the clustering. After defining the regions that are to be aggregated, the network is adjusted using the copper plate approach within each obtained cluster (see chapter Copperplate Aggregation). The capacity expansion model is described in chapter Capacity expansion problem.

Results of the presented methods are divided into two main chapters: In an a-priori analysis we show resulting regions obtained from the presented clustering algorithms before solving the optimisation problem (chapter Evaluation of the Regions). Thereafter we show the convergence of each method in a capacity expansion brownfield approach under a 60% and 100% CO_2 emission cap (chapter Evaluation of the Capacity Expansion model).

At the end, we draw conclusions in chapter Conclusions.

A visualisation of the outline is provided in Fig. 1 using the abbreviations of Table 1 where we additionally outline the novelty of every proposed method.

Methods

Notation, data and model set-up

This study is performed using the open Energy System Model PyPSA-EUR, which is explained in detail in its original publication (Hörsch et al. 2018), where also a partial

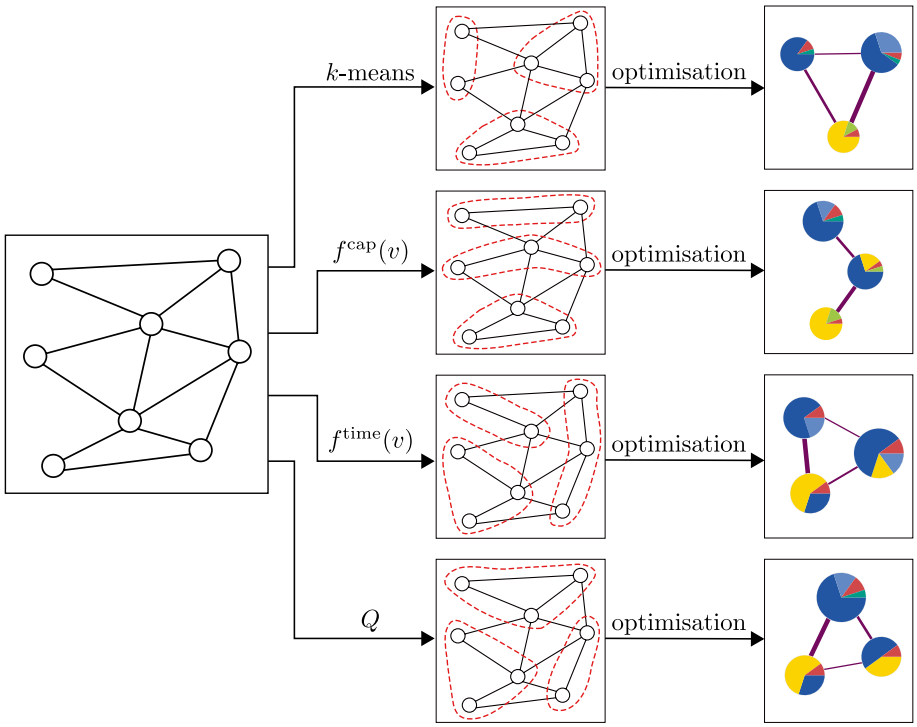

Fig. 1 Graphical representation of the workflow (left to right). An exemplary initial network graph $\mathcal{G}$ is displayed to the left. We apply four aggregation methods (*k*-means, $f^{\text{cap}}(v)$, $f^{\text{time}}(v)$, Q) that are introduced in chapter Methods and summarised in Table 1. Each of the methods considers a different feature of the network. The clustered graphs are used to solve a reduced capacity expansion problem that minimises the total system costs. Results are evaluated on the resulting regions (chapter Evaluation of the Regions)) and on the optimal capacities and power flows (chapter Evaluation of the Capacity Expansion model)

Table 1 Abbreviations and novelty declaration for the applied clustering methods. Each is discussed in an own chapter in the methods chapter, see Methods

abbrev.	Aggregation based on...	Novelty
'country-zones'	... political borders and synchronous zones.	benchmark (no novelty). The spatial resolution of 37 nodes is not variable and the lower bound for all other presented methods.
k-means	... geographic locations (coordinates) of graph nodes. Formulated in eq. (2)	Pre-Aggregation to substations (Dijkstra); otherwise benchmark (no novelty)
$f^{\text{cap}}(v)$	... annual capacity factors of nodes. Formulated in eq. (3) and (4). Hierarchical clustering.	Pre-Aggregation to substations (Dijkstra); thereafter similar to (Siala and Mahfouz 2019) with the following differences: considers network topology using HAC, simultaneous consideration of wind and solar capacity factors in each aggregation step, varying spatial resolution (27 nodes in Siala and Mahfouz (2019))
$f^{\text{time}}(v)$	... hourly capacity factors (time-series) of nodes. Formulated in eq. (3) and (5). Hierarchical clustering.	Fully novel in the context of ESM.
Q	... electrical distance between two nodes. Formulated in eq. (3) and (6). Hierarchical clustering.	Pre-Aggregation to substations (Dijkstra); thereafter similar to Biener and Garcia Rosas (2020), with the difference of accounting for both reactive and resistive parts of transmission lines and considering whole Europe not only Germany to make results comparable.

evaluation of the model is provided. Further validation on curtailment of renewables in the model against historical data was carried out in Frysztacki and Brown (2020) for the years –. There it was shown that the model could portray line congestion accurately. The model contains all existing high-voltage alternating and direct current (HVAC/DC) lines in the European system, as well as those planned by the European Network of Transmission System Operators for Electricity (ENTSO-E) in the Ten Year Network Development Plan (TYNDP). The network topology and electrical parameters of the transmission lines are derived from the ENTSO-E interactive map (2020) using an extraction toolkit (Wiegmans 2016). In the latest release, the model consists of 5323 nodes, 6572 HVAC and 68 HVDC lines (Fig. 2).

Each node can be interpreted as the vertex v of a graph $\mathcal{G} = (\mathcal{V}, E)$, and each transmission line connecting two nodes v and w as an edge (v, w) of $\mathcal{G}$, where $\mathcal{V}$ is the set of all vertices and E the set of all edges. Each node v has its own characteristic attributes, such as its geographical locations given as latitude $x_v \in \mathbb{R}$ and longitude $y_v \in \mathbb{R}$ or a switch $\text{lv}_v \in \{0, 1\}$ to denote whether it is a substation, i.e. connected to the lower voltage distribution grid. Every node with $\text{lv}_v = 1$ is assigned a temporally resolved electricity demand $d_{v,t} \in \mathbb{R}$ in MWh and generation time series $\bar{g}_{v,s,t} \in [0, 1]$ for its renewable carriers $s \in \{\text{solar, onshore wind, offshore wind}\}$. The total electricity demand d_t is taken per country from the Open Power System Data project (2019) and spatially resolved proportional to local population and gross domestic product. Zhou and Bialek (2005) has shown on a sample region in Italy, that this heuristic provides a good correlation. The generation time series are derived using historical wind and solar irradiation data from the ERA5 reanalysis (2020) and the SARAH2 surface radiation dataset (Pfeifroth et al. 2017). Renewable installation potentials $G^{\max}_{v,s} \in [0, \infty)$ are given in MW and are based

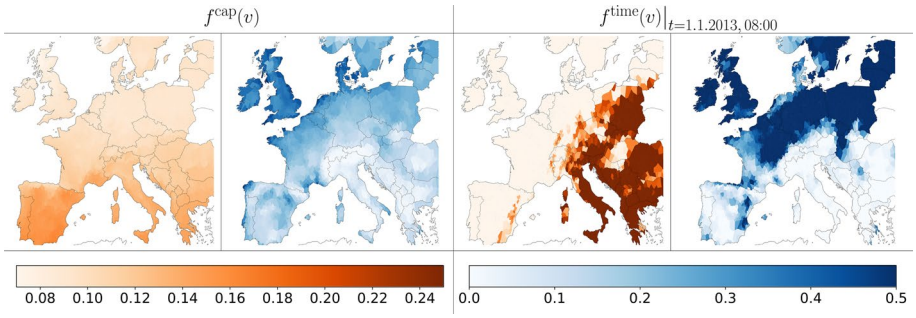

Fig. 2 Examples of the different features for solar (orange) and wind (blue). The left column displays the annual capacity factors, the right column a snapshot at 8 o'clock in the morning on the 1st of January 2013. Both are displayed for the full resolution model of 2435 nodes, recognisable by the white boundaries of the voronoi regions

on land cover maps, excluding for example nature reserves, cities or streets using the geospatial land availability toolkit (Ryberg et al. 2018).

Important attributes of the lines (edges) are their individual resistance $r_{(v,w)} \in \mathbb{R}_0^+$ and reactance $x_{(v,w)} \in \mathbb{R}_0^+$ (both given in Ω), their transmission capacity $F_{(v,w)} \in \mathbb{R}_0^+$ (given in MW) and their length $l_{(v,w)}$ in km.

Clustering methodology

Pre-aggregation

Before applying a clustering method on the model, several preparation steps are conducted to simplify the process. First, all lines are mapped to the voltage level of $380kV$, the prevalent level of the European transmission system.

Second, all one-valent nodes are aggregated to their unique neighbors. This has only a weak effect on renewable generators because of the small cluster sizes, and power flows are not affected strongly because there is only one way for the power to flow from or to one-valent nodes.

In a final pre-aggregation step, a shortest-path problem is solved using Dijkstra's algorithm $D((\mathcal{V}, E), l : E \to \mathbb{R}_0^+)$ on the nodes that are not substations (i.e. $\text{lv}_v = 0$, see Notation, Data and Model Set-Up). Such nodes have no electricity demand, storage units or generators attached. Hence, the same amount of power that flows into such node has to flow out as well, due to the fact that no power can be absorbed or generated. Therefore, neither the power flows nor the generating assets are affected significantly when aggregating them to their electrically closest substations.

All these initial steps reduce the network by approximately a factor of 2 to 2435 nodes, 3673 HVAC and 42 HVDC lines. To further reduce the network size down to a desired number of clusters $37 \leq K \leq 2435$, a clustering method is applied. The lower bound represents the 37 countries and synchronous zones covered by the model. We therefore divide K into 37 integer summands K_z, each representing the number of nodes within a unique associated synchronous zone or country. The clustering methods are respectively applied within each "country-zone" z. The magnitude of K_z is proportional to the electricity demand d_z, for every z:

$$\text{argmin}_{\{K_z\}\in\mathbb{N}^{37}} \sum_{z=1}^{37} \left(K_z - \frac{d_z}{\sum_{z=1}^{37} d_z} K \right) \quad \text{s.t.} \quad \sum_{z=1}^{37} K_z = K\,. \tag{1}$$

The lowest model resolution of 37 nodes represents the benchmark clustering method where every political region is represented by one single node, regardless of the applied clustering method. Thus, each of the methods has the same properties at lowest resolution. When increasing the network resolution beyond 37 nodes, model results start to converge towards the solution at full resolution, where all the methods again yield the same solution, because no clustering is applied. At a resolution of 1250 nodes, the solutions of all discussed methods have sufficiently converged and are therefore taken as benchmark to compare the low resolution solutions to. A detailed survey on why 1250 nodes are a sufficient benchmark is conducted in Appendix, see chapter Sufficient benchmark resolution of 1250 nodes.

K-means clustering

K-means is one of the most commonly applied algorithms in cluster analysis, also in the field of energy system modelling to reduce the initial network to a desired size. It finds groups (clusters) with low variance (with respect to a chosen feature) and favors clusters of similar size. The average complexity is $\mathcal{O}(K|\mathcal{V}|i)$, with number of iterations i. In the worst case, $i = 2^{\Omega(\sqrt{|\mathcal{V}|})}$, resulting in a superpolynomial complexity (Arthur and Vassilvitskii 2006).

In our application the clusters are obtained by solving the minimisation problem

$$\min_{\{(x_c,y_c)^T\in\mathbb{R}^2\}} \sum_{c=1}^{K_z} \sum_{v\in N_c} w_v \left\| \begin{pmatrix} x_c \\ y_c \end{pmatrix} - \begin{pmatrix} x_v \\ y_v \end{pmatrix} \right\|_2^2 \qquad \forall z \in \{1,\ldots,37\}\,, \tag{2}$$

where $(x_c, y_c)^T$ is the mean geographical coordinate of each cluster N_c. The original formulation of *k*-means is designed without weighting w_v. However, we choose a weight proportional to nominal power $G_{v,s}$ for conventional generators s and averaged electricity demand $\langle d_{v,t}\rangle_t$. The weight is chosen such that it incorporates an approximation of the transmission system because it represents how the topology of the network was historically planned to connect major generators to major loads:

$$w_v = \frac{\sum_{s\in\mathcal{S}|_{\text{conv.}}} G_{v,s}}{\sum_{s\in\mathcal{S}|_{\text{conv.}}} \sum_{w\in\mathcal{V}} G_{w,s}} + \frac{\langle d_{v,t}\rangle_t}{\sum_{w\in\mathcal{V}} \langle d_{w,t}\rangle_t}\,.$$

One drawback of *k*-means is that it is not possible to enforce a strict connectivity constraint based on the transmission grid. For example, two nodes that are close in space but not electrically connected can be aggregated to a single node, which can have a significant distorting effect on the power flows. Therefore, the other clustering methods are of hierarchical nature because hierarchical clustering incorporates a connectivity constraint while clustering based a given feature of the data.

Ward's method

Hierarchical agglomerative clustering (HAC) is a bottom-up approach, initially treating each node as a singleton cluster. In each iteration two adjacent clusters are aggregated that have the most similar feature(s) with respect to a given similarity measure. Then, the aggregated cluster's feature is updated. HAC has greedy characteristics, as after the aggregation of the best suited clusters the decision is permanent, and has a running time of $\mathcal{O}(|\mathcal{V}|^2 \log^2 |\mathcal{V}|)$ (Eppstein 2001).

As a distance measure we invoke a variance-minimising approach, similar to k-means. Thus, the distance $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_0^+$ between two clusters N_c and N_d states how much the sum of squares will increase when merging:

$$d(N_c, N_d) = \frac{|N_c||N_d|}{|N_c| + |N_d|} \|\bar{N_c} - \bar{N_d}\|_2^2 \quad \text{where} \quad \bar{N_c} = \frac{1}{|N_c|} \sum_{v \in N_c} f(v)$$

with $f : \mathcal{V} \to \mathbb{R}^d$ being the feature of a node that can be of arbitrary dimension d. This choice of similarity measure is also known as *Ward's method* (Joe and Ward 1963). Recall that initially each node is treated as a single cluster, hence in the first iteration the distance between two nodes is

$$\frac{1}{2} \|f(v) - f(w)\|^2 . \tag{3}$$

In this work, we consider two related, yet different features of the network: The renewable annual capacity factors $\bar{g}_{v,s}$ and the time-series $\bar{g}_{v,s,t}$ of each node, that we briefly present in the following chapters.

Capacity Factor Aggregation

The annual capacity factor $\bar{g}_{v,s} \in [0, 1]$ is a unit-less ratio of the average actual energy output of an asset over its nominal capacity, i.e.

$$\bar{g}_{v,s} = \frac{\langle g_{v,s,t} \rangle_t}{G_{v,s}} ,$$

where $g_{v,s,t}$ is the energy dispatch of asset s in node v at time t. The average is taken over one year, as the name already suggests.

The factors are derived from historical weather data, taking into account the solar irradiation and the wind speeds as well as technical properties of the assets, such as the orientation of solar panels (here: south orientation, tilt angle 35°) or the hub height (here: 80m) of the wind turbine. The capacity factors for wind are obtained from the ERA5 dataset with a spatial resolution of $0.281° \times 0.281°$ (2020), and for solar from the SARAH-2 dataset (Pfeifroth et al. 2017), with a spatial resolution of $0.05° \times 0.05°$. The final capacity factors are derived from the area (excluding the one that is reserved for woodlands, rivers, streets etc.) that is closest to a node v (i.e. the voronoi region) by fully exploiting the available space and placing wind turbines and solar panels. The capacity factors for each location are taken from the characteristic power curves of the assets and then averaged for the corresponding voronoi region.

For Ward's method, we define the feature f in this case as

$$f(v) := f^{\text{cap}}(v) = \bar{g}_{v,s\in\{\text{solar, wind}\}} = \begin{pmatrix} \bar{g}_{v,\text{solar}} \\ \bar{g}_{v,\text{wind}} \end{pmatrix} \in [0,1]^2 . \tag{4}$$

Time Series Aggregation

Resolved capacity factors in time form a series, in this case with a two-hourly resolution over an historical weather year. Without averaging the feed-in over the year, the variability of renewables is evident. For example, the energy production of a solar panel at night is typically zero, while during day time the power output is positive. While the annual capacity factor averages this fact and remains strictly positive for every region, a highly resolved time series captures fluctuations. Thus, additionally to a north-south gradient of the annual capacity factor for solar (higher irradiation in the south) an east-west gradient can be captured (day-night variation). In general, it holds $\bar{g}_{v,s} = \langle \bar{g}_{v,s,t} \rangle_t \, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}$.

The feature f for Ward's method in this case is of high dimension, as every resolved time step has to be considered:

$$f(v) := f^{\text{time}}(v) = \bar{g}_{v,s\in\{\text{solar, wind}\},\, t\in\mathcal{T}} = \begin{pmatrix} \bar{g}_{v,\text{solar},1} \\ \cdots \\ \bar{g}_{v,\text{solar},|\mathcal{T}|} \\ \bar{g}_{v,\text{wind},1} \\ \cdots \\ \bar{g}_{v,\text{wind},|\mathcal{T}|} \end{pmatrix} \in [0,1]^{2|\mathcal{T}|} . \tag{5}$$

$\mathcal{T}$ is the set of all time-steps of the model. In our study, $|\mathcal{T}| = \frac{1}{2} \cdot 8760$, because we resolve our model with a temporal resolution of two hours and run the optimisation over one year (2013).

It is no curse of dimensionality to apply $f^{\text{time}}(v)$, because we solely measure the (high-dimensional) distance between two points; but we do not sample from this high-dimensional space to approximate it with insufficiently many data points.

Clauset-Newman-Moore Greedy modularity maximization

The Clauset-Newman-Moore greedy modularity maximation approach aims to find community structures in large networks. It is a HAC method with approximately linear running time, $\mathcal{O}(|\mathcal{V}|\log^2|\mathcal{V}|)$ (Clauset et al. 2004). In each iteration, it greedily aggregates the two nodes v and w that increase modularity Q the most and continues to do so until the desired number of clusters is reached or until Q can not be further improved.

Q is defined as

$$Q = \frac{1}{2m} \sum_{v,w} \left(A_{vw} - \frac{k_v k_w}{2m} \right) \delta(c_v, c_w) , \tag{6}$$

where A_{vw} is the weighted adjacency matrix of the network graph $\mathcal{G}$, m the sum of all edge weights in the graph, and k_v the weighted degree of node v. These quantities are formally defined as

$$A_{vw} := \begin{cases} w_{(v,w)} & \text{if } (v,w) \in E \\ 0 & \text{otherwise} \end{cases} , \qquad m := \frac{1}{2} \sum_{v,w} A_{vw}, \qquad k_v := \sum_w A_{vw} .$$

The Kronecker-Delta function is given as $\delta(c_v, c_w) := \begin{cases} 1 & \text{if} \quad c_v = c_w \\ 0 & \text{otherwise} \end{cases}$. c_v denotes the cluster node v is assigned to. This means, that the sum in Q is only non-zero, if v and w belong to the same cluster. In its original publication (Clauset et al. 2004), modularity it was introduced without weights, i.e. $w_{(v,w)} = 1$, but we choose a different weighting to adapt the method better to the network topology, similar to the suggestion in Biener and Garcia Rosas (2020), but accounting for both the reactive and resistive components of the grid. We choose the absolute value of the admittance $|y_{(v,w)}|$ of each line (v, w), a measure of electrical distance that describes how easily a circuit allows power to flow. Admittance is defined as the inverse impedance $y_{(v,w)} = \frac{1}{z_{(v,w)}}$.

Regarding the values of Q, A_{vw} is large and positive for a good division, i.e. when aggregating electrically close nodes v and w, and small or zero for a bad division, i.e. when the impedance is high, or if the nodes are not connected at all. $\frac{k_v k_w}{2m}$ is a measure of (electrical) centrality: it tells us, how well the nodes v and w are interconnected in the graph, independent of each other. If the value is large, v and w are nodes with either many connections or they are connected by lines with low impedance. A small value indicates a sparse connection, i.e. either few edges or connections with high impedance. Thus, a (large) positive value of the difference $A_{vw} - \frac{k_v k_w}{2m}$ marks v and w to be electrically closer than they are on average from other nodes in the network. Their aggregation therefore suggests a good grouping. An example is discussed in Fig. 3.

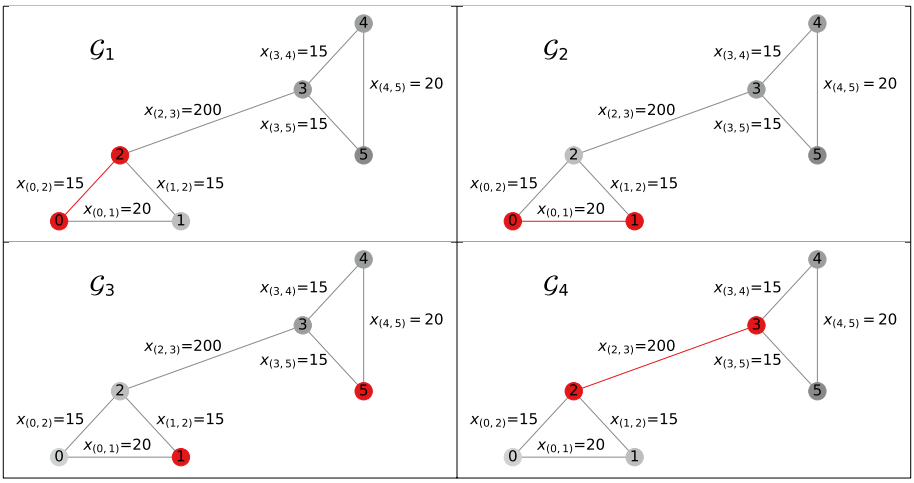

Fig. 3 Consider a symmetric graph $\mathcal{G}_0$ with reactances $x_{(v,w)}$ and without resistances. Above figures show different first iteration choices of the weighted Clauset-Newman-Moore Algorithm into graphs $\mathcal{G}_i$, $i \in \{1, 2, 3, 4\}$ marked in red. Due to the symmetry of G_0, other than the displayed choices for the fist iteration are equivalent. Without clustering, each node in G_0 can be interpreted as a singleton cluster, yielding the initial modularity of $Q_0 \approx -0.1677$. For the four displayed cases, we calculate:

$$\begin{array}{l|l} \mathbf{G_1 : A_{02} \approx 0.067 > \frac{d_0 d_2}{2m} \approx 0.022} & G_2 : A_{01} = 0.05 > \frac{d_0 d_1}{2m} \approx 0.018 \\ \hline G_3 : A_{15} = 0 < \frac{d_1 d_5}{2m} \approx 0.314 & G_4 : A_{23} = 0.005 < \frac{d_2 d_3}{2m} \approx 0.372 \end{array}$$

Hence, both G_1 and G_2 would improve the modularity, but G_1 is the better choice, as $A_{02} - \frac{d_0 d_2}{2m} > A_{01} - \frac{d_0 d_1}{2m}$. G_3 and G_4 are bad choices, reducing modularity and deteriorate the network community. However, if $x_{(2,3)}$ was much smaller, for example $x_{(2,3)} = 1$, then G_4 would be the best choice for the first iteration

Overview of clustering algorithms

In the following chapters, we use the abbreviations introduced in the Methods chapter. We summarise them in Table 1.

Copperplate aggregation

After mapping every node v to a cluster N_c, i.e. $v \mapsto N_c$, all nodes within N_c are replaced by a single equivalent node, where the attributes of all nodes within N_c are aggregated to one equivalent. For example, demand and generation potentials are summed up, and capacity factors are averaged. This replacement is referred to as copperplate approach because it is equivalent to all nodes inside N_c being connected to a lossless copper plate. Finally, all lines (v, w) that connect nodes within the same cluster, i.e. $v, w \in N_c$, are removed from the model, while lines connecting nodes in different clusters, i.e. $v \in N_c \wedge w \in N_d$ where $c \neq d$, are aggregated to an equivalent line.

Capacity expansion problem

The optimisation problem minimises yearly total system costs, including all annualised investment costs $c_{v,s}$ and operation costs $o_{v,s}$. Cost assumptions are based on projections for the year 2030 and derived according to suggestions from the Danish Energy Agency (Technology data for generation of electricity and district heating, energy storage and energy carrier generation and conversion 2019) (wind), the German Institute for Economic Research (Schröder 2013) (conventional technologies, pumped hydro storage, hydro, run-of-river), Budischak et al. (2013) (storage) and the European Technology and Innovation Platform for Photovoltaics Vartiainen et al. (2017) (solar). 2030 is chosen for the cost projections since this is the earliest possible time that such a system transformation might be feasible, and because the cost assumptions are conservative compared to projections for a later year.

The objective function is

$$\min_{\substack{G_{v,s},\\ g_{v,s,t},\\ f_{(v,w),t}}} \left[\sum_{v\in\mathcal{V}}\sum_{s\in\mathcal{S}}\left(c_{v,s}G_{v,s} + 2\sum_{t\in\mathcal{T}} o_{v,s}g_{v,s,t}\right)\right], \tag{7}$$

where $\mathcal{S}$ is the set of all the technologies available for the optimisation. It contains solar, wind both on- and offshore, run-of-river, oil, gas turbines, coal, lignite, geothermal and biomass in terms of generation, and hydrogen, battery and hydro-dams in terms of storage technologies. Nuclear is excluded due to its low social acceptance. The factor of 2 accounts for the 2-hourly resolution in time.

The dispatch of generators $g_{v,s,t}$ has to be non-negative and is constrained by its capacity $G_{v,s}$ multiplied with an hourly capacity factor $\bar{g}_{v,s,t} \in [0, 1]$, that was introduced in chapter Time Series Aggregation. For conventional technologies, $\bar{g}_{v,s,t} = 1$:

$$0 \leq g_{v,s,t} \leq \bar{g}_{v,s,t} G_{v,s} \qquad \forall v \in \mathcal{V}, \forall t \in \mathcal{T}, \forall s \in \mathcal{S}\,. \tag{8}$$

The installable renewable capacity $G_{v,s}$ is bounded below by todays installed capacities $G^{2018}_{v,s}$, and bounded above by land eligibility. The upper bound is derived using the

GLAES tool (Ryberg et al. 2018) and is always finite for renewable carriers. Expansion of conventional generators is not allowed.

The energy levels of all storage units have to be consistent between all hours, accounting for the standing loss, charging efficiency, discharging efficiency, inflow (e.g. river inflow in a reservoir) and spillage. Additionally, the energy level is assumed to be cyclic, i.e. $e_{i,s,t}\big|_{t=0} = e_{i,s,t}\big|_{t=|\mathcal{T}|}$ and is limited by the storage energy capacity $G_{v,s}$.

CO_2 emissions are limited by a cap CAP_{CO2}, implemented using the specific emissions e_s in CO_2-tonne-per-MWh of the fuel s and the efficiency $\eta_{v,s}$ of the generator. In all simulations this cap was set at a reduction of 60% or 100% of the electricity sector emissions compared to 1990.

The (perfectly inelastic) electricity demand $d_{v,t}$ at each node v must be met at each time t by either local generators and storage or by the flow $f_{(v,w),t}$ of a line connected to v. This is required according to Kirchhoff's Current Law (KCL).

In the present paper the linear load flow is used, which has been shown to be a good approximation for a well-compensated transmission network (Stott et al. 2009), including simulations using a large-scale European transmission model (Brown et al. 2016). To guarantee the physicality of the network flows, in addition to KCL, Kirchhoff's Voltage Law (KVL) must be enforced in each connected network. KVL states that the voltage differences around any closed cycle in the network must sum to zero.

The power flows $f_{(v,w),t}$ are also constrained by 70% of their respective line capacities $F_{(v,w)}$

$$f_{(v,w),t} \leq 0.7 \cdot F_{(v,w)}^{\mathrm{TYNDP\,2018}} \,. \tag{9}$$

They are fixed for the optimisation and portray the grid topology that is planned in the TYNDP (2018). The factor of 70% leaves a buffer of 30% of the line capacities to account for $n-1$ line outages and reactive power flows. The choice of 70% is standard in the grid modelling literature (Brown et al. 2016) and is also the target fraction of cross-border capacity that should be available for cross-border trading in the European Union (EU) by 2025, as set in the 2019 EU Electricity Market Regulation (2019).

We perform a brownfield capacity optimisation that builds on a system that exists as of 2018 for both the generating fleet according to the dataset provided in (Open Power System Data 2020) and the planned transmission grid in the ten year network development plan 2018 (TYNDP 2018). The optimisation is subject to two decarbonisation goals of 60% and 100% lower emissions compared to 1990. Missing capacities of renewables for the system to be feasible with respect to the decarbonisation goals are optimised in the sense that the total system costs are minimised.

Results

Evaluation of the regions

First of all we present resulting regions in Fig. 4 for an exemplary spatial resolution of 67 nodes. Additionally, the ranges of cluster sizes are shown in Fig. 5, displaying how many nodes were aggregated into one cluster for varying numbers of clusters in steps of 30. Results on the community structure, i.e. modularity Q given in equation (6), are shown in Fig. 6 for all possible model resolutions starting at 37 and up to 2435 nodes.

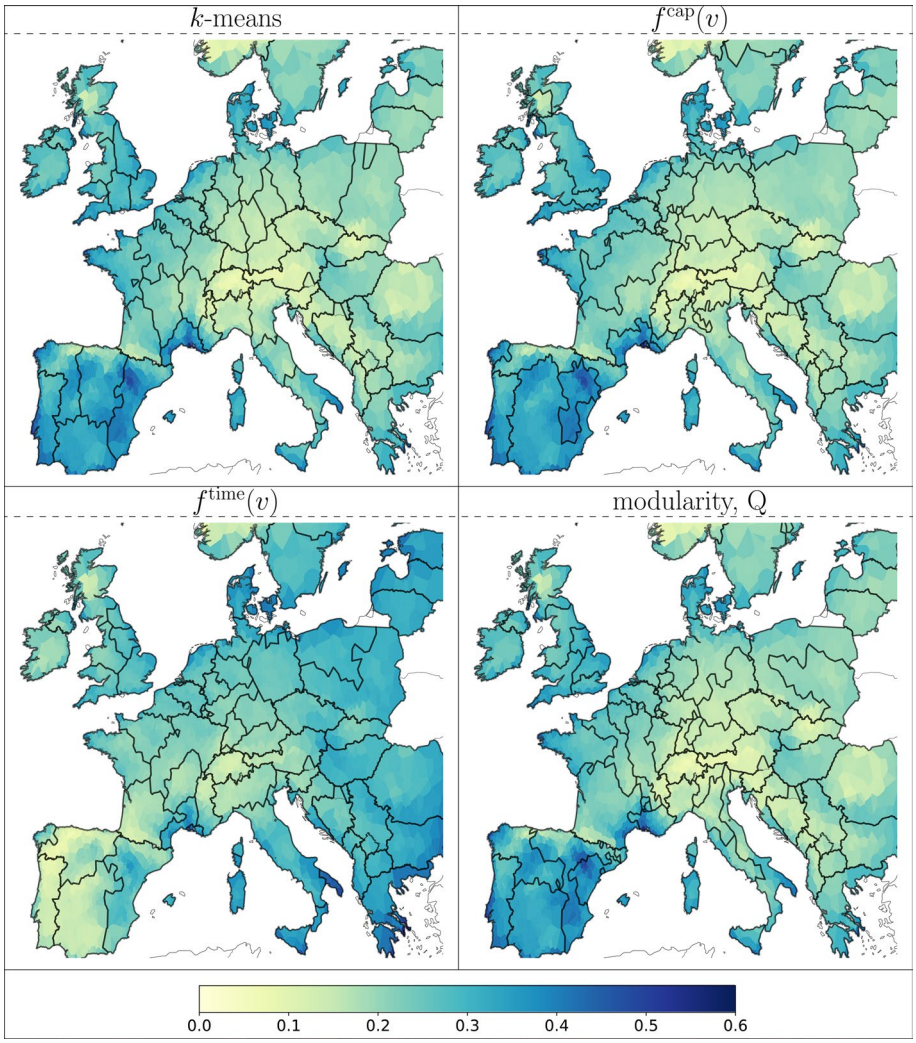

Fig. 4 Resulting regions respective clustering method at a resolution of 67 nodes. The color-map reflects the annual capacity factors for all the methods, except for $f^{\text{time}}(v)$, where it is the average capacity factor of the time-series at 8 o'clock in the moring. Original regions/nodes are highlighted by white boundaries of respective voronoi regions

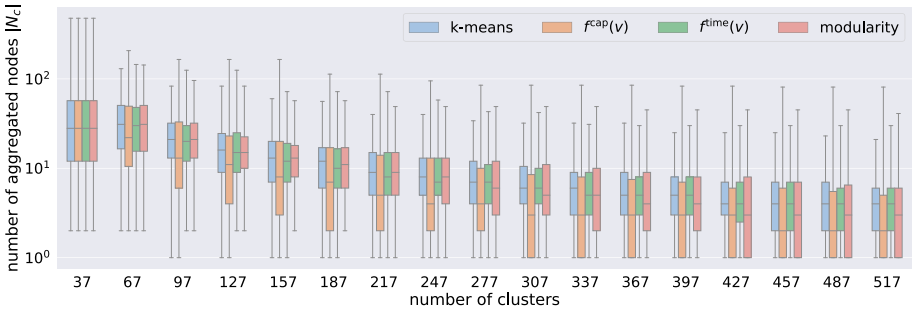

Fig. 5 Clustersizes respective clustering method: The x-axis displays the number of resulting regions, the y-axis the number of aggregated nodes per region, i.e. $|N_c|$. The horizontal line within each bar denotes is the median, the expansion of the bars the 25% and the 75% quantile. The black vertical lines mark the 1.5-times interquartile range

Fig. 6 Modularity respective clustering method: The x-axis displays the number of resulting regions (clusters), the y-axis the modularity of the resulting graph according to equation (6)

Capacity factors are evaluated in Fig. 7 on a quantile base, because the optimisation problem will place renewable assets at their best available sites whenever possible as more power can be generated there with the same cost penalty in the objective function, according to constraint (8). We also present the average full load hours of the renewable assets installed by 2018 in Fig. 8:

$$\sum_{t\in\mathcal{T}} \left\langle \bar{g}_{v,s,t} \Big|_{G^{2018}_{v,s}>0} \right\rangle_v \qquad s \in \{\text{solar, wind}\} \tag{10}$$

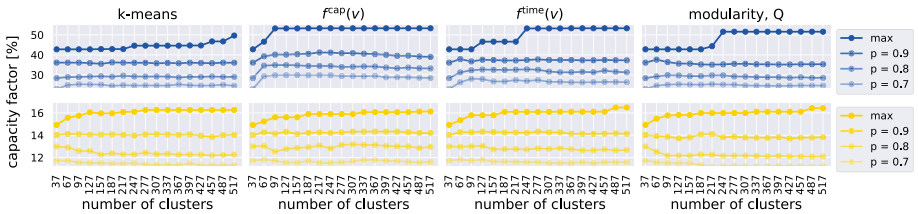

Fig. 7 Quantiles of capacity factors respective clustering method: The x-axis displays the number of resulting regions, the y-axis the resulting 90%, 80% and 70% quantiles of the capacity factors for wind (1st row) and solar (2nd row). 'max' denotes the larges of all available capacity factors

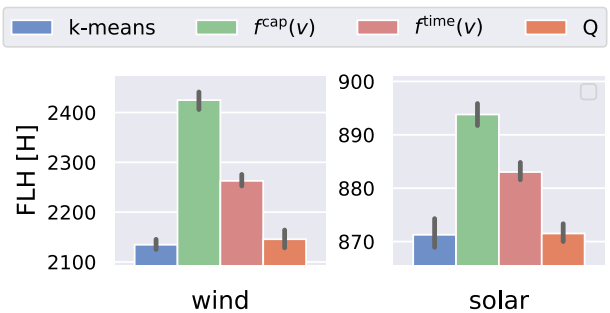

Fig. 8 Average full load hours (FLH) according to equation (10) of existing assets respective renewable technology for all model resolutions except for 37 nodes, because the FLH are equal at this resolution

Evaluation of the capacity expansion model

The main objective of applying different clustering techniques to the model is to reduce the model size for it to be computationally tractable. But at the same time, we want to obtain good estimates for all the optimisation variables introduced in chapter Capacity expansion problem, especially those of equations (7) and (9). It is desired that the low resolution results (estimates) resemble the high resolution model results. For the power flow this means that the sum of flows f of high resolution lines (v, w) that are aggregated to one line (c, d) in the low resolution model (estimate $\hat{f}$) is the same:

$$\hat{f}_{(c,d),t} \stackrel{!}{=} f_{(c,d),t} := \sum_{\substack{(v,w) \in E : \\ v \in N_c \ \wedge \ w \in N_d}} f_{(v,w),t} \qquad \begin{matrix} \forall t \in \{0, \dots, |\mathcal{T}|\} \\ \forall c, d \in \{1, \dots, K\},\ c \neq d \end{matrix} \tag{11}$$

Similarly for the generation and storage capacities. The sum of optimised capacities G at nodes within a cluster $v \in N_c$ should equal the one at the clustered node c at the low resolution model (estimate $\hat{G}$):

$$\hat{G}_{c,s} \stackrel{!}{=} G_{c,s} := \sum_{v \in N_c} G_{v,s} \qquad \begin{matrix} \forall s \in \{\text{solar, wind}\} \\ \forall c \in \{1, \dots, K\} \end{matrix} \quad . \tag{12}$$

The same is desired for the dispatch (or charging/discharging) of all generating (or storing) assets. But we restrict the discussion to those two samples of optimisation variables, as there exists a strong correlation between generation and capacity. This is because exploiting installed resources whenever possible is cheapest according to constraint (8) due to the low operational costs for renewables, $o_{v,s} \approx 0$.

As the model cannot be solved at full resolution for any of the clustering methods, the high-resolution optimised capacities $G_{v,s}$ and power flows $f_{(v,w),t}$ are taken from a model resolution of 1250 nodes (see chapter Sufficient benchmark resolution of 1250 nodes in Appendix for a justification why this benchmark resolution is sufficient), and the estimator quantities $\hat{G}_{c,s}$ and $\hat{f}_{(v,w),t}$ from a model with 97 nodes, the same resolution as in Fig. 7 for the capacity factors. Analysing model results at the spatial resolution of 97 is because many studies choose a resolution of approximately 100 nodes for their research, such as the final report of the e-Highway 2050 project (2015). The mappings of optimal capacities in equation (12) and power flows in equation (11) are shown in Figs. 9 and 10.

Finally, Figs. 11 and 12 display the resulting objective of the optimisation in equation (7) for the two considered CO_2 reduction targets of 60% and 100% for different model resolutions in steps of 30 nodes up to a model resolution of 397 nodes.

Discussion of the results

Discussion on the resulting clusters

In Fig. 4 it can be seen that k-means clustering and Ward's method with hourly capacity factors ($f^{\text{time}}(v)$) favor regions with similar size. For k-means this results from the objective to minimise geographical distances, see eq. (2), while for $f^{\text{time}}(v)$ the reason

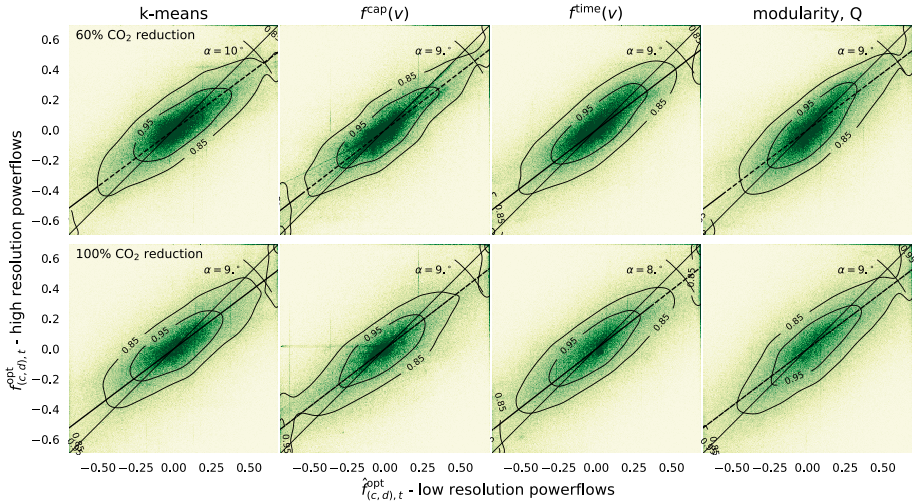

Fig. 9 Normalised mapping of optimal power flows according to equation (11) with the estimated flows of the low resolution model $\hat{f}_{(c,d),t}$ on the x-axis and the aggregated flows on the y-axis. The first row shows results on the 60%, the second row on the 100% reduction target. Instead of presenting the raw data, we plot a two dimensional histogram and outline the 95% and 85% percentiles of the corresponding probability density function (PDF) using contour plots. The black line depicts the origin (perfect correlation), the dashed one a line with the slope of the bivariate correlation coefficient ρ. α is the angle between the two lines

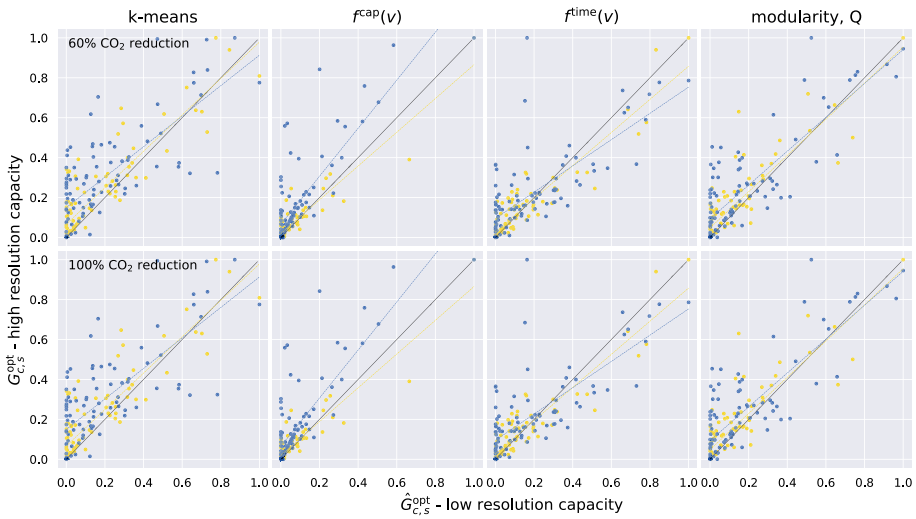

Fig. 10 Normalised mapping of optimal capacities according to equation (12): The x-axis reflects the estimated optimal capacity of a the low resolution model $\hat{G}_{c,s}$, while the y-axis displays the totalised optimal capacities of the high-resolution model. The two considered technologies wind and solar are outlined using different colors. A linear fit to the respective data is added to the plots. A black line depicts a theoretical perfect match

are spatially and temporally varying features that favour this outcome: There exists a north-south gradient for the solar capacity factors that constrains the clusters vertically, and the day-night variation of solar irradiation prevents clusters from being elongated from east to west by adding a high penalty in eq. (3) when trying to merge "day"-nodes with "night"-nodes. This is visible in the east-west elongated clusters as well as large coastal regions that can be observed for $f^{cap}(v)$ because the annual

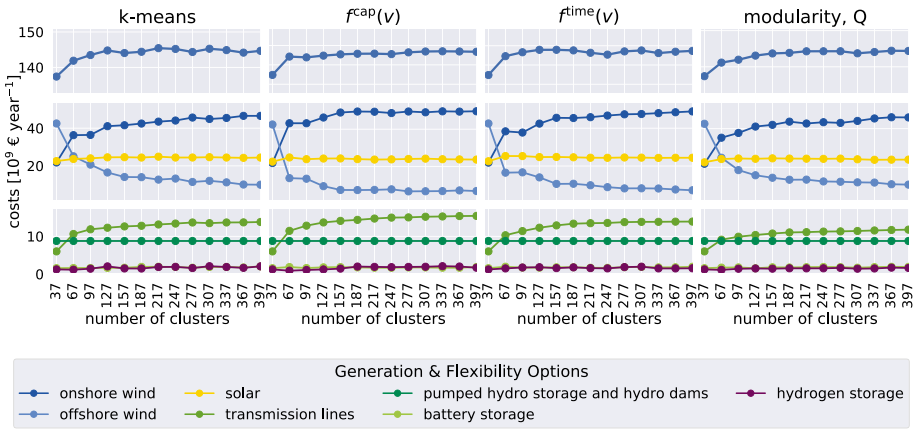

Fig. 11 Resulting total system costs for the 60%-CO_2-reduction scenario according to the objective function (7) respective clustering method. Resolutions in steps of 30 are on the x-axis, investment and operational costs on the y-axis in billion euros. The first row shows the total costs, a breakdown into generation and storage technology is displayed in the second and third row

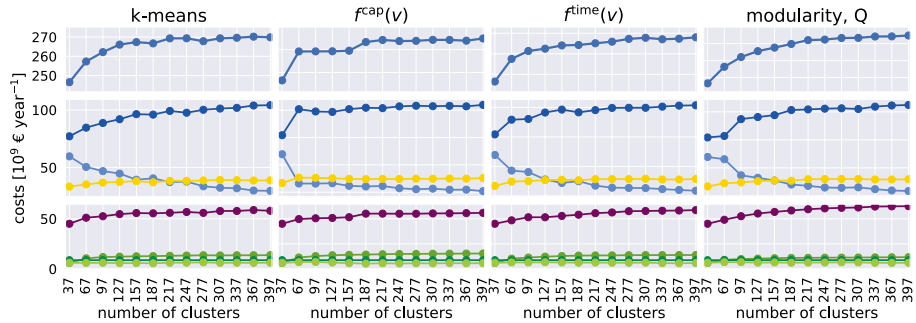

Fig. 12 Resulting total system costs for the 100%-CO_2-reduction scenario

capacity factors do not see these temporal variations. The cluster structure related to modularity Q is similar to $f^{\text{cap}}(v)$, resulting in some very small or elongated thin clusters. The structure of clusters continues so for higher resolutions, as can be seen in Fig. 5. The outcome of long thin clusters is common for single-linkage HAC methods (Everitt et al. 2011), but can be overcome with a profound choice of feature, as we demonstrate using $f^{\text{time}}(v)$.

In Fig. 7 it can be seen that every of the three presented HAC techniques with different similarity/distance measures can capture annual capacity factors better than k-means clustering with respect to the same model resolution. Applying HAC with the similarity measure $f = f^{\text{cap}}(v)$ finds and maintains the best generation sites with an annual capacity factor of 53.3% for wind at a model resolution of 97 nodes. The competing clustering techniques are behind: The best generation site is reached at a model resolution of 247 when invoking hourly capacity factors $f^{\text{time}}(v)$ or modularity Q as a similarity or distance measure and 517 nodes for k-means. For Q as well as for k-means, the best generation site has a lower annual capacity factor of only 51.6% and 49.7% respectively. However, when siting solar assets, the behavior is different: The best site is available earliest for $f = f^{\text{time}}(v)$ and $f = Q$ for a model resolution of 487 nodes and capacity factors

of 16.48% and 16.42%. Both k-means and $f = f^{\text{cap}}(v)$ perform worse, with lower capacity factors even at a model resolution of 512 nodes. This reflects also in the full load hours of existing assets $G^{2018}_{v,s}$ of the respective clustering methods (Fig. 8).

Regarding the community structure of the resulting reduced graph, only HAC based on modularity performs significantly better than the competing methods. Although Ward's method takes into account the structure of the transmission grid by considering only adjacent neighbors, both algorithms perform slightly worse in terms of community structure than k-means, see Fig. 6. Regardless of the method, when reducing the model resolution below a threshold of approximately 40 nodes, modularity suddenly drops to zero.

Discussion on the optimised model results

First, we consider the power flow estimates introduced in equation (12). As we expect the low-resolution flow $\hat{f}_{(c,d),t}$ to equal the aggregated optimal flow $f_{(c,d),t}$, the associated random variable should be distributed proportional to a two dimensional normal distribution:

$$\begin{pmatrix} \hat{f}_{(c,d),t} \\ f_{(c,d),t} \end{pmatrix} \sim \mathcal{N}_2(\mu, \Sigma)\,,$$

where the covariance matrix Σ, and in particular the confidence ellipses provide insight of the correlation between the estimated and the optimal power flow. The length of the axes of the ellipses can be derived form the eigenvalues σ_i of the covariance matrix Σ, namely $r_i = \sqrt{\sigma_i}$. This approximately corresponds to the 40^{th} percentile, i.e. 40% of the data points lie within the ellipse and 60% outside of it. The narrower the minor axis r_2 is, the more data points are close to the origin. The major axis r_1 gives insight of the magnitude of the power flows. The larger the major axis, the more large power flows can be observed. Information on the relation between the actual power flow f and its estimate $\hat{f}$ can be gained from from the bivariate correlation coefficient ρ

$$\rho = \frac{\Sigma_{01}}{\Sigma_{00}\Sigma_{11}} = \frac{COV(\hat{f}_{(c,d),t}, f_{(c,d),t})}{\mathbb{V}(\hat{f}_{(c,d),t}) \cdot \mathbb{V}(f_{(c,d),t})}\,. \tag{13}$$

It is a measure of correlation between the two variables. It is 1 for a perfect correlation, 0 for no correlation and -1 for a negative correlation. Resulting correlation factors ρ and minor axes r_2 can be taken from Table 2.

Sor the 60% carbon reduction target, $f^{\text{cap}}(v)$ as a similarity measure for Ward's method yields the best correlation factor ρ, but the features of annual capacity factors $f^{\text{time}}(v)$ and modularity Q deviate from $f^{\text{cap}}(v)$ by only 0.26% and 1.56% respectively in terms of ρ. The minor axis of the confidence ellipses is also most narrow for the features $f^{\text{time}}(v)$ and $f^{\text{cap}}(v)$, but only 2.5% wider for Q. The distribution of k-means has an approximately 3% lower correlation factor of 0.746 and an 3.13% wider spread in terms of the minor axis of the confidence ellipse compared to $f^{\text{cap}}(v)$ and $f^{\text{time}}(v)$. These variations are clearly visible in Fig. 9. With a higher carbon reduction target of 100%, the trend that clustering on siting capacity prevails over electrical distance when considering the power flow estimates. $f^{\text{time}}(v)$ yields a 3% better correlation ρ and $f^{\text{cap}}(v)$ a 1.45% better one than Q. In terms

Table 2 Bivariate correlation factor ρ and radius of the minor axis r_n of the PDF of power flows in Fig. 9 according to equation (11) for each respective clustering method and carbon reduction target for a spatial resolution of 97 nodes

CO_2Reduction	60%		100%	
	ρ	r_2	ρ	r_2
k-means	0.746	0.165	0.755	0.175
$f^{\text{cap}}(v)$	0.769	0.160	0.768	0.173
$f^{\text{time}}(v)$	0.767	0.160	0.781	0.169
Q	0.757	0.164	0.757	0.179

Table 3 Mean squared error presented as a sum of over- and underestimated optimal estimates ($MSE = MSE^+ + MSE^-$) according to equation (14) respective clustering method, renewable technology and carbon reduction target for a spatial resolution of 97 nodes. Graphically presented in Fig. 10

CO_2Reduction	60%		100%	
Technology	Wind	Solar	Wind	Solar
k-means	0.37 + 3.82	0.01 + 2.80	0.51 + 3.33	0.12 + 1.23
$f^{\text{cap}}(v)$	0.21 + 0.60	0.03 + 1.00	0.01 + 2.22	0.11 + 0.15
$f^{\text{time}}(v)$	0.04 + 3.17	0.08 + 0.79	0.55 + 1.94	0.26 + 0.28
Q	0.36 + 1.31	0.47 + 1.17	0.25 + 1.98	0.17 + 0.78

of the spread (r_2) the same can be observed: The distribution of power flows of $f^{\text{time}}(v)$ is 5.59% narrower than for Q and f^{cap}s distribution is 3.35% slimer than the one of Q. *k*-means performs similar as in the 60% reduction target. To make sure these results are not artificial for the resolution of 97 nodes, we provide the same Table for a spatial resolution of 67 and 127 nodes in Appendix, see chapter More Comparison Results. These results are in-line with the ones we found for a resolution of 97 nodes, where it becomes even more evident that the error made by *k*-means is much larger than the one made by the competing methods.

Considering the mapping of optimal capacity according to equation (12), we could pursue the same approach as for the power flows, i.e. assuming a normal distribution, but due to the relatively low amount of data points, such an analysis would be inaccurate. Instead, we provide the mean squared errors in Table 3:

$$MSE = \frac{1}{K}\sum_{c=1}^{K}\left(G_{c,s} - \hat{G}_{c,s}\right)^2.$$

We distinguish between the over- and underestimated optimal capacities to be able to make better judgement which clustering method is more conservative than another; i.e.

$$MSE = \underbrace{\left(\frac{1}{|\mathcal{K}^+|}\sum_{c\in\mathcal{K}^+}\left(G_{c,s} - \hat{G}_{c,s}\right)^2\right)}_{MSE^+\text{(overestimated capacities)}} + \underbrace{\left(\frac{1}{|\mathcal{K}^-|}\sum_{c\in\mathcal{K}^-}\left(G_{c,s} - \hat{G}_{c,s}\right)^2\right)}_{MSE^-\text{(underestimated capacities)}}, \tag{14}$$

where $\mathcal{K}^+ = \{c \in \{1,\dots,K\} \text{ s.t. } \hat{G}_{c,s} > G_{c,s}\}$ is the set of clusters where optimal capacities are overestimated and analogously $\mathcal{K}^- = \{c \in \{1,\dots,K\} \text{ s.t. } \hat{G}_{c,s} < G_{c,s}\}$ is the set of clusters where optimal capacities are underestimated.

While the clustered models tend to underestimate the need of renewable generation and storage capacity ($MSE^+ \ll MSE^-$) for any of the clustering methods, according to the resulting values presented in Table 3 and Fig. 10, clustering based on f^{cap} performs best in the optimal placement of simultaneously placing wind and solar assets for every carbon reduction target. However, the methods of f^{time} and Q are not significantly worse and yield errors in the same order of magnitude. On the other hand, k-means performs significantly worse with an at least $0.21 - 2.39$ times higher MSE^- value compared to the competing methods. To make sure these results are not artificial for the resolution of 97 nodes, we provide the same Table for a spatial resolution of 67 and 127 nodes in Appendix, see chapter More Comparison Results. These results are in-line with the ones we found for a resolution of 97 nodes.

In terms of storage technologies, no clear tendency can be derived. All methods equally under- and overestimate the need for storage technology ($\mathcal{O}(MSE^+) \approx \mathcal{O}(MSE^-)$) and all clustering methods perform equally well. Values for the MSE can be found in Table 8 in Appendix ("MSE values for Storage" section).

Regarding the total system costs presented in Figs. 11 (60% reduction of carbon emissions) and 12 (100% reduction), we can observe substantially different convergence behaviors of the investment in different technologies and the total system costs. In all of the applied methods a big swing from offshore wind at low resolution of 37 nodes ('country-zones') to onshore wind can be observed, and all methods yield similar results in terms of generation capacity at a spatial resolution of approximately 320 nodes. Ward's method applied with $f = f^{\text{cap}}(\nu)$ converges fastest, where the total costs don't change substantially after reaching a model resolution of 157 nodes (60% reduction) and 67 nodes (100% reduction). At the side of flexibility options, the results need higher spatial resolution than provided to reach an equilibrium as they deviate from one another even at the highest spatial resolution. For the 60% reduction target, the investment in transmission lines is highest for $f^{\text{cap}}(\nu)$ and almost 25% cheaper for Q, because the assets are sourced more locally where demand is high, not exploiting the good sites as they are not available for this clustering. This can be taken from Fig. 10. The same trend continues for the 100% reduction target, but here, for Q, it is clearly visible that 12% more hydrogen storage is needed compared to $f^{\text{cap}}(\nu)$, 8% more compared to k-means and 6% more compared to $f^{\text{time}}(\nu)$. This is because the transmission bottlenecks are better portrayed in Q than in the competing clustering techniques, while the good generation sites are not available to cover demand. This reflects well with Fig. 9.

Limitations of this work

Comparing modeling results retrieved from varying spatial resolutions is a computationally challenging task, meaning that additional simplifications had to be made to the model. For example, the optimisation is run for a single weather year, only those technologies that are considered most substantial in the energy transition are included in the model and the scope of the model is limited to the electricity system. The latter lacks the coupling of different sectors such as building heating, transport and non-electric

industry demand, but including them might offer additional flexibility and interactions and change the results substantially. Nevertheless, a lot of research is conducted based on electricity-only models, such that our results are still valuable. A follow-up study could consider the interactions of spatial scale under different clustering methods in sector-coupled systems.

Regarding our results on the network representation, we ignore the positive impacts that dynamic line rating could impose on the ampacity of the overhead transmission grid. In our simulations, we model severe high summer weather conditions such that the results are conservative However, the ampacity of lines can be significantly increased, which might impair on our results conducted on the electrical distance of the network, where the cooling effects of wind are not considered in the metric (6).

On the other hand, in terms of modeling renewables and particularly offshore wind, we did not model wake effects of wind turbines such that capacity factors for offshore wind are being overestimated. This might impact the strong preference towards offshore wind, particularly for models at low spatial resolution (see Figs. 11 and 12).

Finally, allowing grid-expansion relaxes many of the constraints imposed by the upper bound of line-capacities (9), which in turn will effect the results of this study. However, as found in Frysztacki et al. (2021), grid expansion does not affect the main qualitative features of the results, but it does have the overall effect of lowering the total system costs. Nevertheless, this study could be expanded upon which clustering method captures most of the congested lines and performs best in a planning transmission expansion study.

Conclusions

From this analysis several conclusions can be drawn. First of all, the choice of spatial resolution is crucial to obtain accurate model results, particularly to the ratio and distribution of renewable carriers. A model that is based on political borders such as countries is not advisable, as important transmission bottlenecks are neglected and good generation sites of onshore carriers are underestimated. When moving towards a higher spatial resolution where each country is represented by multiple nodes, modelers should consider carefully how the aggregation is conducted. For models that consist of conventional carriers (such as coal or lignite), an accurate estimate of power flows is more important than accurately portraying renewable generation sites. Therefore, in this case we suggest a model reduction based on electrical distance such as the Clauset-Newman-Moore greedy modularity maximization. However, modeling is mostly conducted to simulate future green scenarios that have high shares of renewable energy. In this case, Ward's method applied on the full time series prevails in terms of accurate siting of capacity and in terms of a good approximation of power flows. It is advisable not to choose annual capacity factors because it ignores correlations in time and leads to very elongated clusters. This tends to underestimate transmission bottlenecks within regions and, therefore, underestimates the need of renewable capacity. Inter-regional power flows in a model where Ward's method based on the full time series was applied for equivalencing are similarly well estimated as those obtained from a reduced model based on electrical distance. For higher shares of renewables the power flow approximation of the reduced model using Ward's method on the time-series is even more precise compared to results

obtained from the reduced model based on electrical distance. Therefore, when modeling a highly renewable electricity system, we recommend using a hierarchical method with a similarity measure that entails spatio-temporal features of renewables, such as the renewable time-series. Model results obtained from clustering on the geographical locations of the nodes are less accurate than those from any of the three hierarchical methods both in terms of siting renewable capacities and an accurate estimate of power flows, so we advise against using this method in future.

Appendix

Sufficient benchmark resolution of 1250 nodes

Computational model feasibility remains a problem even after applying linearisation to the model formulation and spatial/temporal aggregation. Therefore all low-resoltuion model results were compared against a higher resolved model with a spatial resolution of 1250 nodes. Requirements for solving this model size were a runtime of up to 24 days and 240 GB of RAM capacity.

1250 nodes portray approximately 51% of the pre-aggregated model size and 24% of the original full-resolution problem. Results in Frysztacki and Brown (2020) indicate that model results are stable when the spatial resolution is at least 49% of the pre-aggregated model and 26% of the original model. At this or higher resolutions only minor fluctuations of $4-6\%$ occur in terms of optimal dispatch and power flows. As this could potentially be wrong for a capacity expansion problem, we make further justifications for this benchmark by providing the average deviation from the mean as well as the correlation factors for every considered carrier (solar, wind, battery and hydrogen) evaluated on the lowest common region size and power flows between these regions in a 4×4 correlation-matrix. The lowest common regions turn out to be the countries and synchronous zones, which is in line with the benchmark-setting of equation (1).

For the 60% carbon reduction target, optimal investments have small deviations from the mean of up to 5% for offshore wind. Onshore wind and solar installations are more stable with lower cross-deviations. However, optimal installation for battery storage deviates by more than 10% when comparing the 1250 node results of the clustered model with Q to the other clustered model results. But as battery storage for this carbon level is low in general (only 2% of total installed capacity), the relative deviation gives the wrong impression of having strong impact on the optimal result. These results are graphically illustrated in Additional file 1: Fig. S1 and Additional file 2: Fig. S2. It shall also be noted that shifting capacity from one carrier to another might have only small impacts on the objective function, see (Neumann and Brown 2021).

For the 100% carbon reduction target the worst deviation from the mean can be observed for the optimal investment in offshore wind assets with deviations of up to 8%; other technologies as well as power flows have smaller deviations. These results are graphically illustrated in Additional file 3: Fig. S3 and Additional file 4: Fig. S4.

In both scenarios, the pearson's correlation coefficients are ≈ 1 except for power flows where the coefficients are lower but still > 0.9, indicating a linear correlation between the results.

MSE values for storage

We provide the mean squared error values $MSE = MSE^{+} + MSE^{-}$ for storage technologies for a spatial resolution of 97 nodes in Table 8, and additionally for 67 nodes in Table 9 and 127 nodes in Table 10.

More comparison results

To make sure that the results of Tables 2 and 3 are no artifacts of a resolution of 97 nodes and change substantially when varying the spatial resolution, we additionally provide the equivalent tables for of 67 nodes (Tables 4 and 6) and 127 nodes (Tables 5, 6, 7, 8, 9 and 10).

Table 4 Analogous to Table 2, however with a spatial resolution of 67 nodes

CO_2 Reduction	60%		100%	
	ρ	r_2	ρ	r_2
k-means	0.704	0.188	0.725	0.195
$f^{\text{cap}}(v)$	0.754	0.174	0.759	0.187
$f^{\text{time}}(v)$	0.749	0.173	0.765	0.181
Q	0.739	0.173	0.740	0.187

Table 5 Analogous to Table 2, however with a spatial resolution of 127 nodes

CO_2 Reduction	60%		100%	
	ρ	r_2	ρ	r_2
k-means	0.735	0.164	0.772	0.166
$f^{\text{cap}}(v)$	0.802	0.144	0.786	0.163
$f^{\text{time}}(v)$	0.782	0.147	0.808	0.152
Q	0.789	0.152	0.792	0.165

Table 6 Analogous to Table 3, however with a spatial resolution of 67 nodes

CO_2 Reduction	60%		100%	
Technology	Wind	Solar	Wind	Solar
k-means	0.33 + 2.65	0.01 + 2.34	0.22 + 2.43	0.25 + 0.71
$f^{\text{cap}}(v)$	0.23 + 0.79	0.05 + 0.31	0.14 + 1.12	0.05 + 0.12
$f^{\text{time}}(v)$	0.02 + 2.26	0.07 + 0.99	0.51 + 1.63	0.06 + 0.24
Q	0.42 + 1.45	0.16 + 0.71	0.61 + 1.76	0.07 + 0.48

Table 7 Analogous to Table 3, however with a spatial resolution of 127 nodes

CO_2 Reduction	60%		100%	
Technology	Wind	Solar	Wind	Solar
k-means	0.42 + 5.34	0.06 + 2.17	0.51 + 2.22	0.21 + 1.03
$f^{\text{cap}}(v)$	0.79 + 0.86	0.02 + 0.82	0.2 + 1.14	0.11 + 0.15
$f^{\text{time}}(v)$	0.81 + 2.74	0.02 + 1.45	0.14 + 2.38	0.24 + 0.75
Q	0.36 + 1.31	0.47 + 1.17	0.24 + 2.2	0.36 + 1.07

Table 8 Mean squared error presented as a sum of over- and underestimated optimal estimates ($MSE = MSE^{+} + MSE^{-}$) according to equation (14) respective clustering method, storage technology and carbon reduction target for a spatial resolution of 97 nodes. Graphically presented in Fig. 10

CO_2Reduction	60%		100%	
Technology	**Hydrogen**	**Battery**	**Hydrogen**	**Battery**
k-means	0.62 + 0.28	1.0 + 0.76	0.74 + 1.54	0.28 + 0.32
$f^{\text{cap}}(v)$	1.37 + 0.04	0.41 + 1.39	0.24 + 0.68	2.29 + 0.67
$f^{\text{time}}(v)$	0.64 + 0.51	0.57 + 0.28	0.51 + 2.76	0.99 + 0.37
Q	0.58 + 0.07	0.82 + 1.45	0.08 + 2.8	0.09 + 0.26

Table 9 Analogous to Table 8, however with a spatial resolution of 67 nodes

CO_2 Reduction	60%		100%	
Technology	**Hydrogen**	**Battery**	**Hydrogen**	**Battery**
k-means	0.81 + 0.47	0.34 + 0.26	0.44 + 0.64	0.58 + 0.52
$f^{\text{cap}}(v)$	0.89 + 0.33	1.18 + 1.03	0.13 + 0.86	0.43 + 0.01
$f^{\text{time}}(v)$	0.16 + 0.19	0.46 + 0.26	0.52 + 0.65	0.3 + 0.08
Q	0.0 + 0.1	0.75 + 0.44	0.12 + 0.78	0.02 + 0.19

Table 10 Analogous to Table 8, however with a spatial resolution of 127 nodes

Reduction	60%		100%	
Technology	**Hydrogen**	**Battery**	**Hydrogen**	**Battery**
k-means	0.23 + 1.55	0.62 + 1.09	0.34 + 2.75	0.17 + 0.69
$f^{\text{cap}}(v)$	0.24 + 0.96	0.24 + 1.07	2.1 + 1.29	1.92 + 0.96
$f^{\text{time}}(v)$	0.23 + 0.38	0.27 + 0.61	0.94 + 1.93	0.75 + 0.69
Q	0.94 + 0.06	0.34 + 1.59	0.35 + 1.35	0.52 + 0.67

Abbreviations

ENTSO-E	European Network of Transmission System Operators for Electricity
ESM	Energy System Modelling
EU	European Union
EUR	Europe
HAC	Hierarchical Agglomerative Clustering
HVAC/DC	High Voltage Alternating Current / Direct Current
KCL	Kirchhoff's Current Law
KVL	Kirchhoff's Voltage Law
PyPSA	Python for Power System Analysis
TYNDP	Ten Year Network Development Plan

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42162-022-00187-7.

Additional file 1: Figure S1. Correlation factors for every considered carrier (solar, wind, battery and hydrogen) and power flows between the different model results respective clustering at a spatial resolution of 1250 nodes for a 60% carbon reduction target compared to 1990s level.

Additional file 2: Figure S2. Average deviations from the mean for every considered carrier (solar, wind, battery and hydrogen) and power flows between the different model results respective clustering at a spatial resolution of 1250 nodes for a 60% carbon reduction target compared to 1990s level.

Additional file 3: Figure S3. Correlation factors as provided in Additional file 1: Fig. S1, here provided for a scenario where no carbon emissions are allowed

Additional file 4: Figure S4. Average deviations from the mean for different result out-puts as provided. Additional file 2: Fig. S2, here provided for a scenario where nocarbon emissions are allowed.

Acknowledgements
We thank our external colleagues Fabian Hofmann from the Frankfurt Institute for Advanced Studies and Johannes Hampp from University Gießen as well as our co-workers at the Karlsruhe Institue of Technology Fabian Neumann, Elisabeth Zeyen, Kaleb Phipps and Prof. Veit Hagenmeyer for helpful discussions, suggestions and comments.

Authors' contributions
MF: Conceptualisation, Methodology, Software, Formal Analysis, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization. GR: Methodology, Software, Writing—Review & Editing. TB: Conceptualisation, Writing—Review & Editing, Funding Acquisition. All authors read and approved the final manuscript.

Funding
The authors acknowledge funding from the Helmholtz Association under Grant No. VH-NG-1352. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
All data and code is available at zenodo (2020, 2021), apart from the clustering methods. They will be included in a later release. No materials were used for this study.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 26 August 2021 Accepted: 14 February 2022
Published: 4 May 2022

References

Arthur D, Vassilvitskii S (2006) How Slow is the k-Means Method? In: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, pp. 144–153. https://doi.org/10.1145/1137856.1137880
Biener W, Garcia Rosas KR (2020) Grid reduction for energy system analysis. Electr Power Syst Res 185:106349. https://doi.org/10.1016/j.epsr.2020.106349
Brown T, Schierhorn P-P, Ackermann T (2016) Optimising the European transmission system for 77 electricity by 2030. IET Renew Power Gener 10:3–96. https://doi.org/10.1049/iet-rpg.2015.0135
Brown T, Hörsch J, Schlachtberger D (2018) PyPSA: Python for Power System Analysis. Journal of Open Research Software **6**(4). https://doi.org/10.5334/jors.188.1707.09913
Brown T, Hörsch J, Hofmann F, Neumann F, Smith R, Chloe Schlachtberger D, et al. (2020) PyPSA/PyPSA: PyPSA Version 0.17.1. https://doi.org/10.5281/zenodo.3946413
Budischak C, Sewell D, Thomson H, Mach L, Veron DE, Kempton W (2013) Cost-minimized combinations of wind power, solar power and electrochemical storage, powering the grid up to 9.99% of the time. J Power Sourc 225: 60–74. https://doi.org/10.1016/j.jpowsour.2012.09.054
Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111. https://doi.org/10.1103/PhysRevE.70.066111
Cotilla-Sanchez E, Hines P, Barrows C, Blumsack S, Patel M (2013) Multi-attribute partitioning of power networks based on electrical distance. IEEE Trans Power Syst 28(4):4979–4987. https://doi.org/10.1109/TPWRS.2013.2263886
Eppstein D (2001) Fast hierarchical clustering and other applications of dynamic closest Pairs. ACM J Exp Algorith 5:1. https://doi.org/10.1145/351827.351829
European Centre for Medium-Range Weather Forecasts (ECMWF) (2020) ERA5 Reanalysis. https://software.ecmwf.int/wiki/display/CKB/ERA5+data+documentation
European Network of Transmission System Operators for Electricity (ENTSO-E): Interactive Transmission System Map (2020). https://www.entsoe.eu/data/map/
European Network of Transmission System Operators for Electricity (ENTSO-E) (2018) Ten-Year Network Development Plan (TYNDP) . https://tyndp.entsoe.eu/tyndp2018/
European Union (2019) Regulation (EU) 2019/943 of the European Parliament and of the Council of 5 June 2019 on the internal market for electricity. http://data.europa.eu/eli/reg/2019/943/oj
Everitt BS, Landau S, Leese M, Stahl D (2011) Hierarchical clustering, Agglomerative methods. In: Cluster Analysis, pp. 73–84. Wiley, Chichester, West Sussex, U.K. Chap. 4.2
eHighways 2050 Final Reports (2015) Technical report, ENTSO-E and others
Fleischer CE (2020) Minimising the effects of spatial scale reduction on power system models. Energy Strat Rev 32:100563. https://doi.org/10.1016/j.esr.2020.100563
Frysztacki M, Brown T (2020) Modeling Curtailment in Germany: How Spatial Resolution Impacts Line Congestion. In: 2020 17th International Conference on the European Energy Market (EEM), pp. 1–7. https://doi.org/10.1109/EEM49802.2020.9221886
Frysztacki MM, Hörsch J, Hagenmeyer V, Brown T (2021) The strong effect of network resolution on electricity system models with high shares of wind and solar. Appl Energy 291:116726. https://doi.org/10.1016/j.apenergy.2021.116726
Hörsch J, Hofmann F, Schlachtberger D, Brown T (2018) PyPSA-Eur: an open optimisation model of the European transmission system. Energy Strat Rev 22:207–215. https://doi.org/10.1016/j.esr.2018.08.012
Hörsch J, Neumann F, Hofmann F, Peters J, Unnewehr JF, et al. (2021) PyPSA/pypsa-eur: v0.3.0. https://doi.org/10.5281/zenodo.4309093

Joe H, Ward Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845
Kotzur L, Markewitz P, Robinius M, Stolten D (2018) Impact of different time series aggregation methods on optimal energy system design. Renew Energy 117:474–487. https://doi.org/10.1016/j.renene.2017.10.017
Kotzur L, Nolting L, Hoffmann M, Groß T, Smolenko A, Priesmann J, Büsing H, Beer R, Kullmann F, Singh B, Praktiknjo A (2021) A modeler's guide to handle complexity in energy systems optimization, journal = Advances in Applied Energy 4:100063. https://doi.org/10.1016/j.adapen.2021.100063
Kueppers M, Perau C, Franken M, Heger HJ, Huber M, Metzger M, Niessen S (2020) Data-driven regionalization of decarbonized energy systems for reflecting their changing topologies in planning and optimization. Energies 13:16. Doi: https://doi.org/10.3390/en13164076
Lombardi F, Pickering B, Colombo E, Pfenninger S (2020) Policy decision support for renewables deployment through spatially explicit practically optimal alternatives. Joule 4(10):2185–2207. https://doi.org/10.1016/j.joule.2020.08.002
Neumann F (2021) Costs of regional equity and autarky in a renewable European power system. Energy Strat Rev 35:100652. https://doi.org/10.1016/j.esr.2021.100652
Neumann F, Brown T (2021) The near-optimal feasible space of a renewable power system model. Electric Power Syst Res 190:106690. https://doi.org/10.1016/j.epsr.2020.106690
Open Power System Data (2019) Data Platform, Time Series of Load in Hourly Resolution. https://doi.org/10.25832/timeseries/2019-06-05. http://www.open-power-system-data.org/
Open Power System Data (2020) Data Package Renewable Power Plants. https://doi.org/10.25832/renewablepowerplants/2020-08-25. https://data.open-power-system-data.org/renewable_power_plants/
Perera ATD, Nik VM, Chen D, Scartezzini J-L, Hong T (2020) Nat Energy 5:150–159. https://doi.org/10.1038/s41560-020-0558-0
Pfeifroth U, Kothe S, Müller R, Trentmann J, Hollmann R, Fuchs P, Werscheck M (2017) Surface radiation data set - heliosat (sarah) - edition 2. https://doi.org/10.5676/EUM_SAF_CM/SARAH/V002
Pfenninger S, Hawkes A, Keirstead J (2014) Energy systems modeling for twenty-first century energy challenges. Renew Sustain Energy Rev 33:74–86. https://doi.org/10.1016/j.rser.2014.02.003
Radu D, Dubois A, Berger M, Ernst D (2021) Model Reduction in Capacity Expansion Planning Problems via Renewable Generation Site Selection. In: 2021 IEEE Madrid PowerTech, pp. 1–6. https://doi.org/10.1109/PowerTech46648.2021.9495027
Ryberg D, Robinius M, Stolten D (2018) Evaluating land eligibility constraints of renewable energy sources in Europe. Energies **11:**5. https://doi.org/10.3390/en11051246
Sasse J-P, Trutnevyte E (2020) Regional impacts of electricity system transition in Central Europe until 2035. Nat Commun. https://doi.org/10.1038/s41467-020-18812-y
Schlachtberger DP, Brown T, Schramm S, Greiner M (2017) The benefits of cooperation in a highly renewable European electricity network. Energy 134:469–481. https://doi.org/10.1016/j.energy.2017.06.004
Schröder A, Kunz F, Meiss J, Mendelevitch R, von Hirschhausen C (2013) Current and prospective costs of electricity generation until 2050. Data Documentation, DIW 68, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin. http://hdl.handle.net/10419/80348
Shayesteh E, Hobbs BF, Söder L, Amelin M (2017) ATC-based system reduction for planning power systems with correlated wind and loads. IEEE Trans Power Syst 30(1):429–438. https://doi.org/10.1109/TPWRS.2014.2326615
Shi D, Tylavsky DJ (2015) A novel bus-aggregation-based structure-preserving power system equivalent. IEEE Trans Power Syst 30(4):1977–1986. https://doi.org/10.1109/TPWRS.2014.2359447
Shi D, Shawhan DL, Li N, Tylavsky DJ, Taber JT, Zimmerman RD, Schulze WD (2012) Optimal generation investment planning: Pt. 1: network equivalents. In: 2012 North American Power Symposium (NAPS), pp. 1–6. https://doi.org/10.1109/NAPS.2012.6336375
Siala K, Mahfouz MY (2019) Impact of the choice of regions on energy system models. Energy Strat Rev 25:75–85. https://doi.org/10.1016/j.esr.2019.100362
Stott B, Jardim J, Alsac O (2009) DC power flow revisited. IEEE Trans Power Syst 24(3):1290–1300. https://doi.org/10.1109/TPWRS.2009.2021235
Technology data for generation of electricity and district heating, energy storage and energy carrier generation and conversion. Technical report, Danish Energy Agency and Energinet.dk (2019). https://ens.dk/en/our-services/projections-and-models/technology-data
Tröndle T, Lilliestam J, Marelli S, Pfenninger S (2020) Trade-offs between geographic scale, cost, and infrastructure requirements for fully renewable electricity in Europe. Joule 4(9):1929–1948. https://doi.org/10.1016/j.joule.2020.07.018
UN Treaty Collection (2012) Paris Agreement. UNTC XXVII 7.d
Vartiainen E, Masson G, Breyer C (2017) The True Competitiveness of Solar PV: A European Case Study. Technical report, European Technology and Innovation Platform for Photovoltaics. http://www.etip-pv.eu/fileadmin/Documents/ETIP_PV_Publications_2017-2018/LCOE_Report_March_2017.pdf
Ward JB (1949) Equivalent circuits for power-flow studies. Trans Am Inst Electr Eng 68(1):373–382. https://doi.org/10.1109/T-AIEE.1949.5059947
What Is the European Green Deal? (2019). https://doi.org/10.2775/97540
Wiegmans B (2016). GridKit extract of ENTSO-E interactive map. https://doi.org/10.5281/zenodo.55853
Zeyen E, Hagenmeyer V, Brown T (2020) Mitigating heat demand peaks in buildings in a highly renewable European energy system. Technical report. arXiv:2012.01831
Zhou Q, Bialek JW (2005) Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades. IEEE Trans Power Syst 20(2):782–788. https://doi.org/10.1109/TPWRS.2005.846178

Publisher's Note