Check for updates

METHOD ARTICLE

REVISED A data processing approach with built-in spatial resolution reduction methods to construct energy system models

[version 2; peer review: 2 approved]

Christian Etienne Fleischer

Department of Energy and Environmental Management, Europa-Universität Flensburg, Flensburg, 24943, Germany

v2 **First published:** 19 Apr 2021, **1**:36
https://doi.org/10.12688/openreseurope.13420.1
Latest published: 10 Feb 2022, **1**:36
https://doi.org/10.12688/openreseurope.13420.2

Abstract

Introduction: Data processing is a crucial step in energy system modelling which prepares input data from various sources into a format needed to formulate a model. Multiple open-source web-hosted databases offer pre-processed input data within the European context. However, the number of documented open-source data processing workflows that allow for the construction of energy system models with specified spatial resolution reduction methods is still limited.
Methods: The first step of the data-processing method builds a dataset using web-hosted pre-processed data and open-source software. The second step aggregates the dataset using a specified spatial aggregation method. The spatially aggregated dataset is used as input data to construct sector-coupled energy system models.
Results: To demonstrate the application of the data processing process, three power and heat optimisation models of Germany were constructed using the proposed data processing approach. Significant variation in generation, transmission and storage capacity of electricity were observed between the optimisation results of the energy system models.
Conclusions: This paper presents a novel data processing approach to construct sector-coupled energy system models with integrated spatial aggregations methods.

Keywords

data processing, energy system modelling, spatial aggregation, sector-coupling

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 10 Feb 2022	✓ view	✓ view
version 1 19 Apr 2021	? view	? view

1. **Tim Tröndle**, ETH Zürich, Zürich, Switzerland

2. **William Usher**, KTH Royal Institute of Technology, Stockholm, Sweden

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Horizon 2020 gateway.



This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.



This article is included in the Energy Systems Modelling collection.

Corresponding author: Christian Etienne Fleischer (christian.fleischer@uni-flensburg.de)

Author roles: Fleischer CE: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No [765515], (project ENSYSTRA).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Fleischer CE. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Fleischer CE. **A data processing approach with built-in spatial resolution reduction methods to construct energy system models [version 2; peer review: 2 approved]** Open Research Europe 2022, **1**:36 https://doi.org/10.12688/openreseurope.13420.2

First published: 19 Apr 2021, **1**:36 https://doi.org/10.12688/openreseurope.13420.1

REVISED **Amendments from Version 1**

As noted in the reviewers' comments, the proposed data processing approach is the strongest part of the article. Therefore, the article type has been changed to a Methods article to focus the article's content around the data processing approach.

The result and discussion section have been adapted to primarily present the case studies built using the data processing approach. Table 2 has been added to the Methods section to present the variables used to calculate the weighting factors during aggregation of the capacity factors. Additionally, a Jupyter notebook was added to the archived source code to demonstrate the different steps of creating the datasets and the case studies presented in the article.

A different data source has been used to increase the spatial resolution of the temperature variable in the Areas dataset. The heat demand and coefficient of performance of air-source heat pumps variables have been updated to sub-national levels using the updated temperature data.

Figure 2 was changed to add the hydrogen storage option in the illustrative description of the power system model. Batteries have been omitted as a storage option in the new case studies to reduce the complexity of the models.

Given the changes made to the data processing approach and the modelling components, new results were generated and represented in the updated Figure 3.

Furthermore, minor concerns such as referencing mistakes and the use of more updated datasets have been addressed in the new version

Any further responses from the reviewers can be found at the end of the article

Introduction

In the past, energy system models were primarily closed and proprietary. However, recently more open-source energy system modelling tools have been made available. Maruf *et al.* identified 59 freely available energy system modelling tools[1]. Energy system models are considered 'open' when the data and model code is accessible and legally usable[2]. Pfenninger *et al.* discuss how open models improve the scientific quality of the models by adhering to fundamental scientific principles such as transparency and reproducibility[3]. Pfenninger *et al.* also state that when models and data are open, productivity increases as it reduces the time spent by researchers in duplication of work in developing models and datasets[3].

The steps in the open-source energy modelling process are described by Pfenninger *et al.* in 2. One crucial step in that process is data processing. Data processing is an intermediate step between the raw input data and the model formulation. The input data is made accessible to the formulated model after undergoing data processing. The methods used to process the input data can have an impact on model results. Two documented impacts are the effects of temporal resolution reduction methods[4–9] and spatial resolution reductions methods[10–14]. Therefore, the data processing steps must be well documented to ensure that their impact on the modelling results can be properly gauged. There are a limited amount of available open-source modelling tools and datasets that allow for the alteration of the spatial resolution of energy system models. One of these tools is presented in 15 by Hörsch *et al.*, which builds a highly spatially disaggregated European power system model dataset. The resolution of the dataset can then be reduced at various spatial scale by clustering the electrical network using the k-means algorithm. Tröndle *et al.* investigate the possibility of renewable energy autarky at various spatial scales using European power system models at four different spatial scales: continental level, national level, regional level and municipal level[16].

Input data of high spatial and temporal resolution can be generated using tools such as the global Renewable Energy atlas (REatlas) atlas[17], the Python Generator of renewable time series and maps (PyGreta)[18], and the GlobalEnergyGIS[19]. The input data can also be obtained from an extensive list of web-hosted platforms, repositories and databases. These platforms and datasets include the renewables.ninja platform[20], the Open Power System Data (OPSD) platform[21], the hotmaps repository[22] and the ENSPRESO database[23]. The Open Energy Platform compliments these platforms by documenting and sharing datasets used by existing energy system models such as the eTraGo[24], OSeMBE[25] and MEDEAS[26]. These platforms facilitate the identification of documented and validated input data in centralised locations.

This paper presents a novel data processing approach that maximises the use of the web-hosted pre-processed input data to build energy system models. The appliction of the data processing approach is demonstrated by building three power and heat models with different spatial contexts. The differences between the spatial contexts are the spatial scope and spatial zones that define the regions in the models.

Methods

Data processing workflow approach

The proposed data processing approach can be split into two steps as illustrated in Figure 1. The first step builds the Areas dataset, which host the necessary data variables from the pre-processed input data sources within a structured framework. A set of requirements are defined for the Areas dataset to ensure standardisation and proper documentation of the data variables. The first requirement prescribes that the data variables need to be indexed using standardised reference keys. These reference keys allow the data to be uniformly organised according to spatial, temporal, and technological specifications. The Nomenclature of Territorial Units for Statistics level 2 (NUTS 2) and NUTS level 0 are the two spatial reference keys used by the Areas dataset to structure data with a spatial dimension. The second requirement ensures the use of standardised units of measurements. The standardisation of units ensures uniformity, allowing the use of the dataset for modelling without additional unit conversions. The final requirement for building an Areas dataset is the documentation of the data variables in the dataset. The documentation entails providing the source of the data variables and a description of the unit.

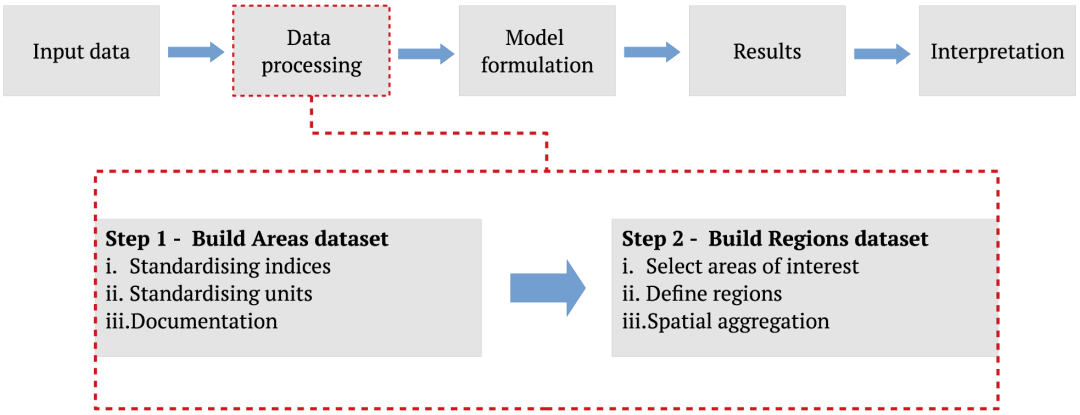


Figure 1. An illustrated description of the proposed data processing workflow of the energy system modelling process. The energy system modelling process is based on the Openmod Philosophy laid out by Pfenninger *et al.* in 2.

Building a Regions dataset is the second step of the data processing process. The Regions dataset is constructed by first defining the areas of interest. The areas of interest could include all areas in the Areas dataset, or it could consist of a subset of areas. As the name of the dataset indicates, the Regions dataset allows for the grouping of areas into regions. The Regions dataset is described in more detail in *Regions dataset*.

Dataset framework. Two requirements were defined to guide the selection process of the dataset framework used as the skeleton for the Areas and the Regions dataset. The first requirement is that the dataset framework should be able to handle data with more than one reference key. The need for multiple reference keys handling capability allows data variables to be referenced according to spatial, temporal, or even technological indices. The second requirement is that the dataset framework must integrate well with existing open-source energy modelling software and scientific analysis software. As there is a multitude of these software[15,27–30] written in the Python programming language[31], the dataset framework should be a Python-based software. The xarray dataset object in the xarray toolkit[32] was selected to build the datasets based on these two requirements.

There are additional benefits of using xarray to construct the datasets, as listed below:

- the dataset can be exported as a unidata network common (nc) file format that can be compressed to lower file sizes which eases sharing of the datasets;
- the dataset framework allows for documentation of the data variables.

The datasets created can be stored and shared using a single file, in online archives such as Zenodo. Zenodo attaches Digital Object Identifiers (DOIs), which allows for the citation of the data. Hörsch *et al.*[15] and Tröndle *et al.*[33] both share different versions of their model on Zenodo using the nc file format.

Areas dataset. The Areas dataset spatial scope includes the EU 27 countries with the exclusion of Cyprus and Malta and the addition of Norway, Great Britain, and Switzerland. The data variables in the dataset can be sub-divided into two sets, the base variables and the derived variables. The base variables are used to determine the derived variables.

The Areas and the Regions datasets have a total of seven reference keys. There are two spatial reference keys, namely *NUTS 0* and *NUTS 2*. Data variables using the *NUTS 0* reference key apply at the national level, whereas the *NUTS 2* level data are, for most countries, spatial administrative areas within the countries. There are four technology-specific reference keys: *techs*, *fuels*, *techs hydro,* and *techs hydro subset*. The *techs* and *fuels* reference keys reference the different technology types of the power plants variable. The *techs hydro* reference key is used to reference the technologies used to generate power from hydropower. The *techs hydro subset* dimension refers to a subset of the *techs hydro* reference key of hydropower technologies that have reservoirs to store water. The *time* reference key is used to index the time dimension of data variables. The *time* indices are the hourly intervals of a single representative year. The variables reference key, unit and sources are summarised in Table 1. Except for temperature and offshore wind capacity factor, the data variables in the Areas dataset are organised at the NUTS 2 spatial level. Therefore, the Areas dataset can be considered as a collection of data variables of 270 NUTS 2 areas.

Base variables

There are a set of base variables needed to establish the foundation of the dataset. The identification code (NUTS 2 id), the geometrical information (Geometry), and the country identification code (Country code) of the NUTS 2 areas are base variables

Table 1. Summary of the data variables in the Areas dataset.

Data variable	Base/Derived variables	Reference key	Unit	Sources
NUTS 2 id	Base	*NUTS 2*	NA	34
Geometry	Base	*NUTS 2*	NA	34
Country code	Base	*NUTS 2*	NA	34
Temperature	Base	*time, NUTS 2*	Degrees Celsius	35
Population	Base	*NUTS 2*	people	36
Power plants	Base	*techs, fuels, NUTS 2*	MW	37–39
Hydropower plants dispatch	Base	*techs hydro, NUTS 2*	MW	40
Hydropower plants storage	Base	*techs hydro subset, NUTS 2*	MW, MWh	40
Rooftop solar PV area	Base	*NUTS 2*	km^2	41–44
Ground-mounted solar PV	Base	*NUTS 2*	km^2	41–44
Onshore wind area	Base	*NUTS 2*	km^2	22,45,46
Offshore wind area	Base	*NUTS 2*	km^2	45,47
Solar capacity factor	Base	*time, NUTS 2*	Per unit capacity	20
Wind capacity factor	Base	*time, NUTS 2*	Per unit capacity	20
Offshore wind capacity factor	Base	*time, NUTS 0*	Per unit capacity	20
Hydropower capacity factor	Base	*time, NUTS 2*	Per unit capacity	48
Power	Base	*time, NUTS 0*	MW	49
Heat	Derived	*time, NUTS 2*	MW	22,35
Air-sourced capacity factor	Derived	*time, NUTS 2*	Per unit capacity	35,50

taken from the Eurostat database[34]. The NUTS 2 geometry scale is 1:10 million. The ambient air temperature obtained from the ERA-NUTS meteorological time-series dataset[35] is a base variable in the NUST 2 area dataset. The population values in the NUTS 2 areas are extracted from the Global Human Settlement (GHS) population grid[36]. The population raster for the year 2015 was used. The population values within the geometry of the NUTS 2 area are summed, and the sum is the population value assigned to the NUTS 2 area. Other base variables describe the energy generation and storage technologies.

The power plants data variable gives the aggregated installed capacity of conventional power plants, solar photovoltaic (PV) installations, onshore and offshore wind installations associated with the NUTS 2 areas. Information on conventional power plants is from the conventional power plants dataset hosted on the OPSD platform[39]. Information on solar and onshore wind installations for NUTS 2 areas in Germany, Denmark, France, Poland, United Kingdom and Switzerland, were extracted from the OPSD renewable power plants dataset[37]. A power plant from the two datasets is associated with a NUTS 2 area when the power plants' geometrical information places it within the geometrical boundaries of that NUTS 2 area. The power plants are grouped by the fuel and technology type and aggregated by their installed generating capacity in megawatts (MW). As the offshore wind installations are not within the geometries of the NUTS 2 areas, they are determined separately. The existing offshore wind installations were extracted from the European Marine Observation and Data Network (EMODnet) offshore wind farm database[38]. This database provides the location of each wind farm as a georeferenced point and references it to a country. The offshore wind farms were assigned to the closest NUTS 2 area of the country it was referenced too.

The dataset has three categories of hydropower technologies: run-of-river hydropower, reservoir-based hydropower and pumped storage hydropower obtained from the Joint Research Council (JRC) hydropower plants database[40]. The NUTS 2 areas were assigned the cumulative installed capacity of the different hydropower capacities. The reservoir-based and pumped storage hydropower plants' cumulative storage capacity within the NUTS 2 areas was also calculated and added to the

dataset. In the instances where the storage capacity was not given, it was assumed that the plant had a reservoir that can store the water needed to operate the plant at nominal capacity for six hours.

The availability of onshore wind and solar were assigned to the dataset as hourly capacity factor values extracted for each NUTS 2 area from renewables.ninja platform[20]. The capacity factors for offshore wind at the national level was taken from the same data platform. The country-wide daily inflow from 48 defines the capacity factor of the hydropower plants. provides the historical daily inflow in 30 European countries between 2003 to 2012. The hydropower plants' capacity factors were calculated by dividing the daily inflow values by the sum of installed hydropower capacity within the country.

The data variables that define the area available for renewable energy technologies are rooftop solar PV area, ground-mounted solar PV area, onshore wind area and offshore wind area. For the NUTS 2 areas in the EU-27 countries and the United Kingdom, the ENSPRESSO database was used to assign the data variables of rooftop and ground-mounted solar PV area[41] and the data variables for onshore and offshore wind area[45]. The areas classified in the EU-wide low restrictions with 400 m setback distance scenario were selected to define the onshore wind areas. The onshore wind areas in NUTS2 areas of Switzerland were calculated from the wind energy potential areas raster provided by the swiss energy ministry[46]. The onshore wind areas in NUTS 2 areas in Norway were calculated from the wind energy potential areas raster of the hotmaps project[22].

The areas classified within the EU-wide low restrictions with water depth 0 – 30 m and water depth 30 – 60 m scenario were selected from 45 to determine the offshore wind area. Except for the NUTS 2 areas in Norway, the portion of the offshore wind areas assigned to a NUTS 2 area is proportional to their share of their respective country's total coastline. The NUTS 2 areas of Norway are assigned offshore wind areas according to their proximity to the offshore wind areas listed in the Norwegian offshore wind strategic environmental assessment report[47].

The rooftop and ground-mounted solar PV areas in Switzerland and Norway were calculated using the Open Street Map building footprint data[44] and literature values. For Switzerland, the rooftop solar PV area in a NUTS 2 area was proportional to their share of the total building footprint area in Switzerland multiplied by a total available rooftop area of 267 km^2 and a rooftop suitability factor of 0.564 provided by Walch *et al.* in 42. For Norway, the rooftop solar PV area was calculated by multiplying the total building footprint area within the NUTS 2 areas with an rooftop area suitability factor of 0.49 calculated by Bódis *et al.* in 43. The ground-mounted solar PV area in Norway and Switzerland was calculated using the ratio of the ground-mounted solar PV area to the rooftop solar PV of Sweden and Austria, respectively. These ratios are 176:1 and 144:1, respectively.

Derived variables

Hourly electrical load profiles for European countries are only available at the country level from the European Network of Transmission System Operators (ENTSO-E) transparency platform[49]. The load profiles are given at NUTS 0 spatial resolution in the Areas dataset. While building the Regions dataset, the spatial resolution of the load profiles is first reduced to NUTS 2 level before they are aggregated to the spatial resolution of the defined region. This process is discussed in more detail in the Regions dataset subsection.

Using a bottom-up approach, the heat demand profiles $D_{a,t}$ are generated for each NUST 2 area a and time step t using the following equation:

$$D_{a,t} = D_a \cdot \sum_{s,e} [\sigma_{a,s,e} \cdot d_{a,s,e,t}] \quad (1)$$

The bottom-up approach classifies the heat demand in two end-use categories e and two sectors s. The end-use categories are space heating and domestic hot water heating. The sectors are the tertiary and domestic sector. Both end-use categories of each sector have a share factor $\sigma_{a,s,e}$ and a normalised hourly profile $d_{a,s,e,t}$ with time steps t. The share factor gives the percentage contribution of an end-use category of a sector to the total space and water heating demand. These share factors are country-specific and are obtained from the hotmaps repository[22]. The hotmaps repository does not provide share factor values for Norway and Switzerland[22]. Therefore, the share factors for Sweden and Luxemburg were used respectively instead. The normalised profiles are generated at the national level using generic profiles for space heating and water heating obtained from the hotmaps repository[22]. The generic profiles for space heating are country, season and temperature-dependent whereas, the generic profiles of hot water heating only vary according to the day of the week and the season. The normalised space heating profiles are defined using the temperature data in the dataset. NUTS 2 areas within the same country are assigned the same normalised space heating demand profile. The heat demand volume d_a for space heating and hot water heating is calculated from a rasterised map generated by the hotmaps project[22]. The map depicts the estimated final energy demand for space and water heating on each hectare for EU28, Norway, Iceland and Switzerland for 2015.

The temperature variables from the dataset are used to calculate the hourly efficiency factors of the heat pumps.

The following quadratic regression equation, presented by Ruhnau *et al.*[50], is used to determine the coefficient of performance $COP_{t,a}$ of the air-sourced heat pumps:

$$COP_{t,a} = 6.08 - 0.09 \cdot \Delta T_{t,a} + 0.0005 \cdot \Delta T_{t,a} \quad (2)$$

Where $\Delta T_{t,a}$ is the temperature difference between the heat sink temperature and the ambient air temperature. The heat sink temperature is assumed to be a constant value of 50°C. As suggested by Ruhnau *et al.* the calculated $COP_{t,a}$ is adjusted for real-work effects using a correction factor of 0.85.

Regions dataset. Spatial resolution reduction is often used to reduce the computational demand of solving energy system optimisation problems. Depending on the research question or study focus, the data can be aggregated into regions to reduce the spatial resolution of the dataset. A common spatial resolution reduction method used by energy system modellers is to aggregate the spatial data according to political or administrative boundaries. European countries are classified according to multiple NUTS levels. The political regions method can thus group areas according to the NUTS level specified. For example, the spatial resolution of the data for Germany would reduce from 38 government regions of the NUTS 2 areas to the 16 states of NUTS 1. The spatial resolution could also be further reduced to a national level by aggregating the NUTS 2 area data to the NUTS 0 level. The number of NUTS areas at different levels is dependent on the European country. There are spatial resolution reduction methods that group areas according to the heterogeneity of spatial attributes of NUTS 2 areas. The max-p regions method for example, presented by Fleischer[12], groups areas into regions that are similar in population; wind and solar resource potential; and pumped-hydro storage capacity. The max-p regions method uses the max-p-regions algorithm, introduced by Duque *et al.* in 51.

Once the regions are defined, the variables of the NUTS 2 areas are aggregated to have variables that represent the regions. The resulting spatially aggregated dataset, hereafter referred to as the Regions dataset, is used to store and organise the variables generated after data aggregation. In the Regions dataset, the *NUTS 2* reference key is replaced by the *Regions* reference key. The *Regions* reference key is composed of the *NUTS 2* reference keys of the NUTS 2 area within the regions created. The geometry of the NUTS 2 areas attributed to the same region are joined to form the geometry of the regions.

As mentioned in the derived variables section, the spatial resolution of the electrical power profile in the Areas dataset is at the NUTS 0 level. Therefore the power profiles need to be disaggregated to NUTS 2 area spatial resolution before they can be aggregated to the specified regions spatial resolution. Population and Gross Domestic Product (GDP) are commonly used as a proxy to determine the distribution of electrical demand[15,52,53]. Robinius *et al.*, presents a method to disaggregate electricity demand at sub-national levels, but as this method is determined using data for Germany, it is not applicable to all European countries. The chosen proxy to disaggregate the hourly load profiles in the presented case studies is population. In this proposed data processing approach, the NUTS 2 areas assume a share of the load profiles of their respective country. The proportion of the share is calculated by multiplying the country level power profile with the NUTS 2 area-specific weighing factor. In the case studies presented in this paper, population is used to calculate the weighting factors used to disaggregate the power profiles. The weighting factor of a NUTS 2 area is the share of the population in that area in relation to the population of the NUTS 2 area respective country. This approach could be improved as higher spatial resolved data for power profiles for European countries become available. The offshore wind capacity factors are also at NUTS 0 spatial resolution, similar to power profiles. Therefore, offshore wind capacity factors are also disaggregated to NUTS 2 spatial resolution before aggregating them to build the Regions dataset. The capacity factors for offshore wind at NUTS 2 are assumed to be the same as the respective country-level capacity factors.

When building the Regions dataset, the capacity factors of the variable renewable technologies are multiplied by a weighting factor before they are summed. The proxy variable used to determine the weighting factor is technology-specific. The weighting factor is the share of the proxy variable relative to the proxy variable's sum within a region. The technologies and their respective proxy variables used to calculate the weighting factors are given in Table 2.

All other variables do not represent mean values and are summed without weighting factors.

Model formulation. There are some additional items needed in conjunction with a Regions dataset to formulate an energy system model. The first item is an energy system framework. There is a selection of open-source energy modelling frameworks that can be used. The selection of the framework depends on the focus of the study and the preference of the modeller. As the Regions database is generated using the python programming language, it can be integrated well into a python-based modelling framework.

Together with some additional items, the Regions dataset can then be used to formulate energy system models. One essential item is the techno-economic parameters. The techno-economic parameters will depend on the scenarios being investigated by the model. The scenarios also dictate certain assumptions used in the model.

Power and heat optimisation model development

The proposed data processing workflow, implemented in the EUropean Sustainable Energy System (EU-SES) modelling tool[54], is used to build power and heat optimisation models to demonstrate the versatility of the data processing approach and the importance of spatial context in energy system modelling. The EU-SES tool uses the calliope framework[27] to

Table 2. Proxy variables used to aggregate the capacity factors in the case studies.

Data variable	Proxy variable
Solar capacity factor	Sum of Rooftop solar PV area and ground-mounted solar PV
Wind capacity factor	Onshore wind area
Offshore wind capacity factor	Offshore wind area
Hydropower capacity factor	Hydropower plants dispatch
Air-sourced capacity factor	Population

formulate the models. The scripts used to generate the datasets, models and the optimisation results of each model can be found Zenodo[54]. The current version of the EU-SES tool can only automate the construction of an energy system model using the calliope framework. However, as the datasets are separated from models, the datasets can be used as input data in other modelling frameworks such as PyPSA.

The first model is a multi-national model containing ten countries in the NUTS 2 area dataset named the GER NUTS0 model. These ten countries include Germany and nine countries that have a transmission connection with Germany. The NUTS 2 areas spatial data are aggregated according to national jurisdiction in the GER NUTS0 model. The second and third model reduces the spatial scope to include only Germany with no energy import or export from neighbouring countries. The difference between the second and third model is the spatial resolution reduction method used. The second model, named the GER NUTS1 model, are defined according to 16 administrative jurisdictions given by the NUTS 1 level. Whereas the regions in the third model, named the GER MAX-P model, are defined using the max-p regions method to generate nine regions. As illustrated in part a) of Figure 3, GER NUTS 1 model has more regions and therefore, the regions have, on average, a higher spatial resolution than the regions in the GER MAX-P model. The regions with the highest spatial resolution in the GER NUTS 1 model represent the city-states of Berlin, Hamburg and Bremen.

The reference year selected to create the Areas dataset is 2011. The structure of the models is illustrated in Figure 2. The examples are modelled using the calliope modelling framework.

The models all share several overarching scenario assumptions. The following key assumptions are made in this scenario:

- The cumulative CO_2 equivalent emission of the optimised model is limited to 20 % of the 1990 CO_2 equivalent emission of the countries in the model;
- The cumulative biogas available to the cogeneration plants is 420 PJ which was estimated to be a projected value for 2020 presented by Scarlat *et al.* in 55;
- All power plant capacities classified as biomass, gas and cogeneration in the Regions dataset are summed under the classification cogeneration;

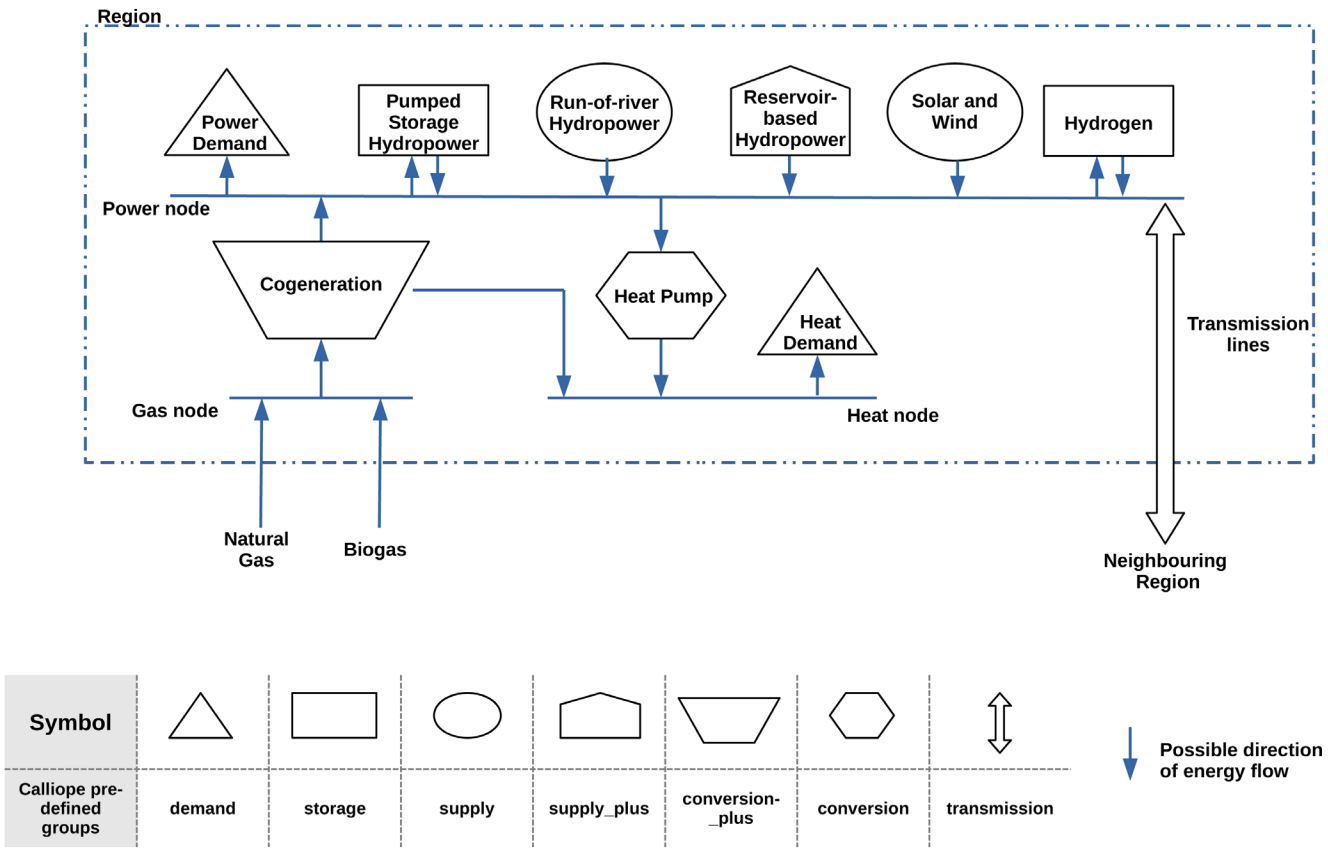


Figure 2. An illustrated description of the model created using the regions dataset and the calliope modelling framework. The table in the figure indicates the predefined technology groups used to describe the different components in the model.

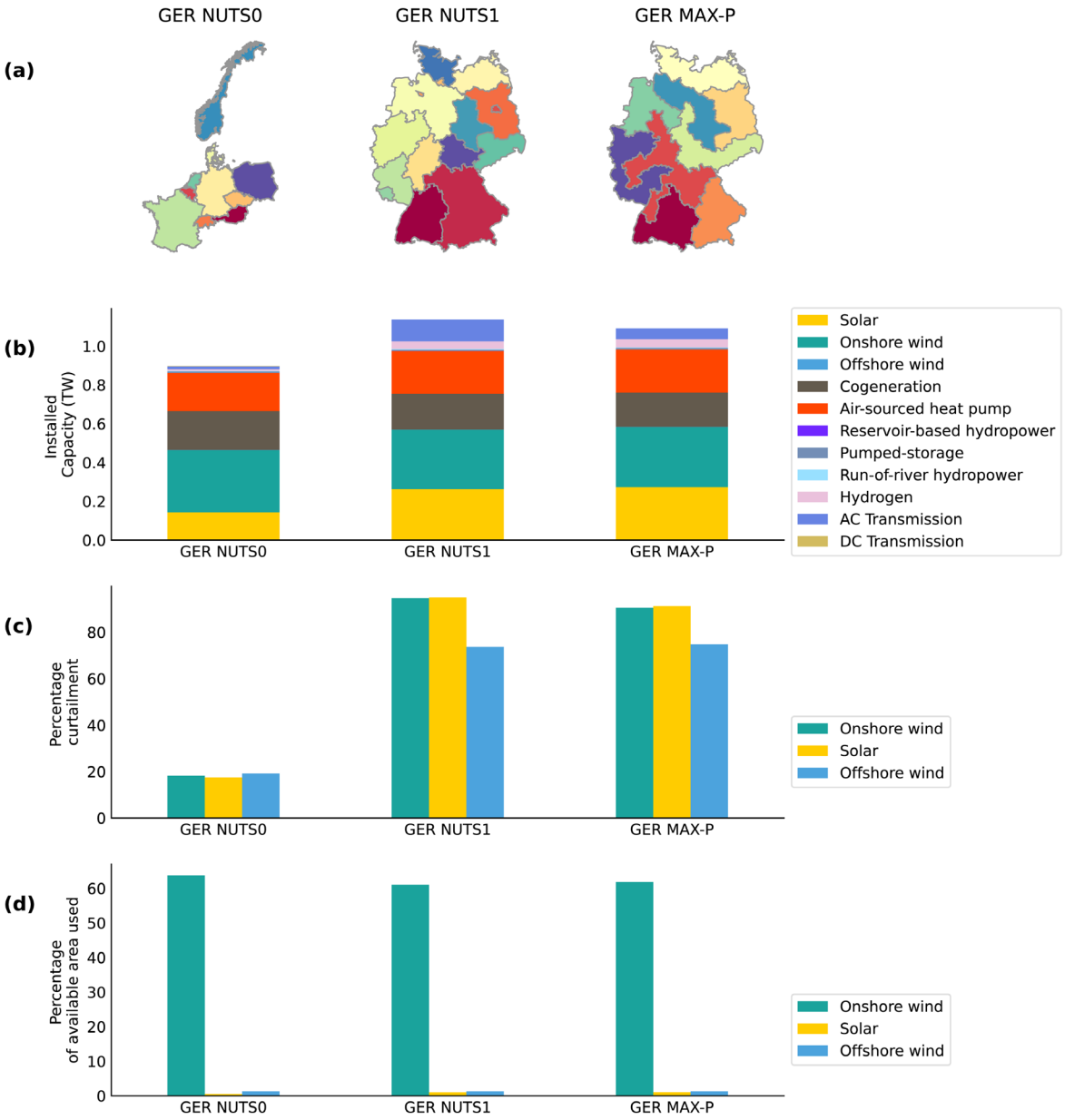


Figure 3. The regions in the GER NUTS0 model, GER NUTS1 model and the GER Max-P model, are illustrated in **a**). The least-cost optimisation results of the three models for Germany are given in **b**), **c**) and **d**). Plot **b**) illustrates the optimised installed capacity of the technologies. The curtailment rate of solar PV, onshore and offshore wind generation is given in percentage value in **c**). Plot **d**) illustrates the optimised percentage of the available area utilised by solar, onshore and offshore wind installations.

- Power plants classified as nuclear, coal, oil, other, waste and geothermal in the Regions dataset are not available in the model;
- The storage level of all storage capacities is assumed to be full during the first and last instance of the optimisation;
- The hydropower power plants, existing installed wind and solar capacities must be adopted in the optimised model;
- The solar and onshore wind capacity density is assumed to be 170 MW/km^2 and 5 MW/km^2, respectively, adopted from Ruiz *et al.*[23];
- Offshore wind installations have a capacity density of 5.36 MW/km^2 adopted from Hundleby and Freeman[56];
- Regions are considered 'copper plates', meaning that within the regions there are no constraints in energy transfer.

The power exchange between regions is possible and is constrained by the net transfer capacity and the efficiency of the power lines. There are two power transfer mediums in the model. The first is the high voltage alternative current (HVAC) transmission lines between regions that share a border. The HVAC has a set rated capacity of 2 GW. The other power transfer

mediums are direct current high voltage interconnectors installed between regions. The list of interconnectors and their respective rated transfer capacity is taken from the installed and planned DC links listed by Fleischer *et al.* in 12. Losses are not considered in the interconnectors and HVAC transmission lines.

The power and heat optimisation model's objective function is to minimise the investment cost and dispatch cost of the model for one year and at a three-hour resolution. The optimisation models assume perfect foresight, and the power and heat demand is inelastic. A discount rate of 7% is assumed to calculate the annualised cost of the investments. The model uses techno-economic parameters projected for the year 2030, documented as *Extended data*[57]. The techno-economic parameters for the generation and storage technologies were adopted from values presented by Moles *et al.*[58] and by Jülch[59], respectively.

The cumulative CO_2 emission constraint ensures that the models have high solar and wind penetration levels. This emission constraint aligns with the roadmap presented in 2011 by the European Commission that aims to reduce 80 % of the EU CO_2 emission by 2050. In 2019 the EU commission revised the CO_2 emission target for 2050 to a net-zero emission target[60]. Therefore the 80 % reduction target could represent a snapshot along the net-zero pathway.

Results and discussions

The optimisation results of the three models are compared in Figure 3. The results in Figure 3 show that the GER NUTS 0 model has the lowest installed capacity of solar PV. This is despite the fact that Germany is represented at a lower resolution in the GER NUTS 0 model than the two other models and does not have the opportunity to maximise the use of good solar sites within Germany. This relatively low solar PV installed capacity of the GER NUTS0 model can be explained by the fact that the GER NUTS0 model has a greater spatial scope that the two other models. This additional benefit in spatial scope allows the GER NUTS0 model to maximise the use of resources available in neighbouring countries to Germany, such as hydropower-based energy storage capacities in Norway, Switzerland and Austria. These storage capacities can help minimise the curtailment rate of the solar PV installations, as illustrated in part c) of Figure 3. The lower-cost hydropower storage capacities in neighbouring countries can also explain why Germany in GER NUTS0 model invest less in expensive hydrogen storage in comparison to the two other models. These apparent differences between the GER NUTS0 model and the models with a different spatial scope document the importance of spatial context in energy system modelling. Part d) of Figure 3 illustrates that more than half of the available onshore area in Germany is used for deploying onshore wind in all three models.

Next, the optimisation results of the two models with the same spatial scope, the GER NUTS 1 model and the GER MAX-P model, are presented and discussed. The optimised transmission capacity of the GER NUTS 1 model is significantly greater than that of the GER MAX-P model, as can be seen in part b) of Figure 3. The fact that the GER NUTS 1 model has more regions, it can have more transmission lines, and therefore it can also have a higher installed transmission capacity value than the GER MAX-P model. As shown in part c) of Figure 3, the optimised GER NUTS 1 model has a slightly higher percentage in curtailment for solar PV and onshore wind, which could be a consequence of more transmission capacity bottlenecks between regions. The differences between the two models that have the same spatial scope but constructed using two different spatial resolution reduction methods demonstrate the importance of spatial context in energy system modelling.

In the following paragraph some reflections are made on the proposed data processing approach. Firstly, this data processing demonstrates that it is possible to automate the construction of sector-coupled energy system models for European countries using exiting web-hosted datasets. Secondly, there are certain data gaps that influence the data processing approach. The first data gap is the lack of power profiles at sub-national spatial resolution. Due to this data gap the power of profiles of a NUTS 2 area is simply assumed to be a portion of the country-level power profile. The portion of the profile is calculated using population data. This method of disaggregating power profiles does not consider certain differentiations between NUTS 2 areas other than population that influence the power profile such as energy intensive industries in areas with low population. To mitigate this limitation this issue certain models also use GDP when disaggregate power profiles. There are also some data gaps in relation to hydropower plants. These data gaps are power plant specific inflow data and storage capacity of power plants with storage reservoirs. Similarly a limited amount of research has been conducted on the impact of spatial aggregation methods on data products particularly on the impact on capacity factors of variable renewable energy technologies and demand profiles.

Conclusion

A novel data processing workflow that maximises the use of the web-hosted validated pre-processed input data to build energy system models is presented. The proposed data processing workflow has a two-step process. The first step organises and standardises the pre-processed input data into a dataset called the Areas dataset. In the second step, the spatial data in the Areas dataset is aggregated according to regions and standardised into a Regions dataset. With the addition of techno-economic parameters and a modelling framework, the Regions dataset can be used to build power and heat models. The data processing approach is not integrated into any specific energy modelling framework, giving the modeller the flexibility to create a power and heat model using the modelling framework best suited for the research question. The proposed approach also provides a baseline that can be extended upon to include other energy sectors such as industry and transport. The proposed workflow is used to build three power and heat optimisation models. The three optimisation models' result demonstrates the importance of how the spatial scope and the method used for spatial resolution reduction can impact the optimisation result.

Data availability

Underlying data

Open Science Framework: A data processing approach with built-in spatial resolution reduction methods to construct energy system models. https://doi.org/10.17605/OSF.IO/JHMXN[57].

This project contains the following underlying data:

- Optimisation results (optimisation results of generated models saved as nc format files)
- Dataset (An areas dataset generated using the proposed data processing saved as nc format file)

Extended data

Open Science Framework: A data processing approach with built-in spatial resolution reduction methods to construct energy system models. https://doi.org/10.17605/OSF.IO/JHMXN[57].

This project contains the following extended data in Extended data figures and tables.docx:

- Appendix 1 – Techno-economic parameters of generation technologies used the example models.
- Appendix 2 – Techno-economic parameters of storage technologies used the example models.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Software availability

- Source code available from: https://github.com/ENSYS-TRA/EU-SES/tree/v1.3
- Archived source code at time of publication: https://doi.org/10.5281/zenodo.5834185[54].
- License: Apache License 2.0 license.

Acknowledgements

The author would like to thank Md. Nasimul Islam Maruf, André Harewood, Abhinav Bhaskar and Annabelle Constance for their comments and suggestions on the paper. The author is thankful to the ENSYSTRA and ZNES team for support and guidance.

References

1. Maruf MNI: **Sector Coupling in the North Sea Region—A Review on the Energy System Modelling Perspective.** *Energies.* 2019; **12**(22): 4298.
Publisher Full Text
2. Pfenninger S, Hirth L, Schlecht I, *et al.*: **Opening the black box of energy modelling: Strategies and lessons learned.** *Energy Strategy Rev.* 2018; **19**: 63–71.
Publisher Full Text
3. Pfenninger S, DeCarolis J, Hirth L, *et al.*: **The importance of open data and software: Is energy research lagging behind?** *Energy Policy.* 2017; **101**: 211–215.
Publisher Full Text
4. Nahmmacher P, Schmid E, Hirth L, *et al.*: **Carpe Diem: A Novel Approach to Select Representative Days for Long-Term Power System Models with High Shares of Renewable Energy Sources.** *SSRN Electron J.* 2014; 28.
Publisher Full Text
5. Pfenninger S: **Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability.** *Appl Energy.* 2017; **197**: 1–13.
Publisher Full Text
6. Poncelet K, Delarue E, Six D, *et al.*: **Impact of the level of temporal and operational detail in energy-system planning models.** *Appl Energy.* 2016; **162**: 631–643.
Publisher Full Text
7. Kotzur L, Markewitz P, Robinius M, *et al.*: **Impact of different time series aggregation methods on optimal energy system design.** *Renew Energy.* 2018; **117**: 474–487.
Publisher Full Text
8. Heuberger CF, Rubin ES, Staffell I, *et al.*: **Power capacity expansion planning considering endogenous technology cost learning.** *Appl Energy.* 2017; **204**: 831–845.
Publisher Full Text
9. Frew BA, Jacobson MZ: **Temporal and spatial tradeoffs in power system modeling with assumptions about storage: An application of the POWER model.** *Energy.* 2016; **117**(1): 198–213.
Publisher Full Text
10. Krishnan V, Cole W: **Evaluating the value of high spatial resolution in national capacity expansion models using ReEDS.** In: *2016 IEEE Power Energy Soc Gen Meet (PESGM).* IEEE, Boston, MA, USA, 2016; 1–5.
Publisher Full Text
11. Hörsch J, Brown T: **The role of spatial scale in joint optimisations of generation and transmission for European highly renewable scenarios.** *2017 14th International Conference on the European Energy Market (EEM).* 2017; 1–7.
Publisher Full Text
12. Fleischer CE: **Minimising the effects of spatial scale reduction on power system models.** *Energy Strategy Rev.* 2020; **32**: 100563.
Publisher Full Text
13. Simoes S, Zeyringer M, Mayr D, *et al.*: **Impact of different levels of geographical disaggregation of wind and PV electricity generation in large energy system models: A case study for Austria.** *Renew Energy.* 2017; **105**: 183–198.
Publisher Full Text
14. Siala K, Mahfouz MY: **Impact of the choice of regions on energy system models.** *Energy Strategy Rev.* 2019; **25**: 75–85.
Publisher Full Text
15. Hörsch J, Hofmann F, Schlachtberger D, *et al.*: **PyPSA-Eur: An Open Optimisation Model of the European Transmission System.** *Energy Strategy Rev.* 2018; **22**: 207–215.
Publisher Full Text
16. Tröndle T, Pfenninger S, Lilliestam J: **Home-made or imported. On the possibility for renewable electricity autarky on all scales in Europe.** *Energy Strategy Rev.* 2019; **26**: 100388.
Publisher Full Text
17. Andresen GB, Søndergaard AA, Greiner M: **Validation of Danish wind time series from a new global renewable energy atlas for energy system analysis.** *Energy.* 2015; **93**: 1074–1088.
Publisher Full Text
18. Siala K, Houmy H: **tum-ens/pyGRETA: python Generator of REnewable Time series and mAps**. 2020.
Publisher Full Text
19. Mattsson N, Verendel V, Hedenus F, *et al.*: **An autopilot for energy models –**

Automatic generation of renewable supply curves, hourly capacity factors and hourly synthetic electricity demand for arbitrary world regions. *Energy Strategy Rev.* 2021; **33**: 100606.
Publisher Full Text

20. Pfenninger S, Staffell I: **Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data.** *Energy.* 2016; **114**: 1251–1265.
Publisher Full Text

21. Wiese F, Schlecht I, Bunke WD, *et al.*: **Open Power System Data – Frictionless data for electricity system modelling.** *Appl Energy.* 2019; **236**: 401–409.
Publisher Full Text

22. Pezzutto S, Zambotti S, Croce S, *et al.*: **D2.3 WP2 Report – Open Data Set for the EU28.** (n.d.) 158.
Reference Source

23. Ruiz P, Nijs W, Tarvydas D, *et al.*: **ENSPRESO - an open, EU-28 wide, transparent and coherent database of wind, solar and biomass energy potentials.** *Energy Strategy Rev.* 2019; **26**: 100379.
Publisher Full Text

24. **openego/eTraGo**. openego, 2020; (accessed July 2, 2020).
Reference Source

25. Henke H: **OSeMBE – An open-source engagement model.** KTH Royal Institute of Technology, 2018; (accessed August 3, 2018).
Reference Source

26. Samsó R, Ollé JS: **Guiding European Policy toward a low-carbon economy. Modelling sustainable Energy system Development under Environmental And Socioeconomic constraints.** n.d. (accessed July 2, 2020).
Reference Source

27. Pfenninger S, Pickering B: **Calliope: a multi-scale energy systems modelling framework.** *J Open Source Softw.* 2018; **3**(29): 825.
Publisher Full Text

28. Dorfner J, Schönleber K, Dorfner M, *et al.*: **tum-ens/urbs: urbs v1.0.1.** *Zenodo.* 2019.
Publisher Full Text

29. Hilpert S, Günther S, Kaldemeyer C, *et al.*: **Addressing Energy System Modelling Challenges: The Contribution of the Open Energy Modelling Framework (oemof).** (n.d.). 2017; 2017020055.
Publisher Full Text

30. Howells M, Rogner H, Strachan N, *et al.*: **OSeMOSYS: The Open Source Energy Modeling System: An introduction to its ethos, structure and development.** *Energy Policy.* 2011; **39**(10): 5850–5870.
Publisher Full Text

31. van Rossum G, Drake FL, Van Rossum G: **The Python language reference, Release 3.0.1 [Repr.].** Python Software Foundation, Hampton, NH. 2010.

32. Hoyer S, Hamman J: **xarray: N-D labeled Arrays and Datasets in Python.** *J Open Res Softw.* 2017; **5**(1): 10.
Publisher Full Text

33. Tröndle T: **timtroendle/possibility-for-electricity-autarky: v1.0.** *Zenodo.* 2019.
Publisher Full Text

34. European Commission: **Eurostat (ESTAT), GISCO, Nomenclature of Territorial Units for Statistics (NUTS) 2013 - Statistical Units - Data set.**2015; (accessed April 25, 2019).
Reference Source

35. De Felice M, Kavvadias K: **ERA-NUTS: meteorological time-series based on C3S ERA5 for European regions (1980–2020)**. 2021.
Publisher Full Text

36. European Commission, Joint Research Centre; Columbia University, Center for International Earth Science Information Network: **GHS-POP R2015A - GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015).** European Commission, Joint Research Centre (JRC) [Dataset] PID: 2015; (accessed March 1, 2021).
Reference Source

37. Schlecht I, Simic M: **Renewable power plants.** 2020.
Publisher Full Text

38. EMODnet: **Emodnet_HA_WindFarms_ 20200305.** (n.d.). (accessed June 24, 2020).
Reference Source

39. Weibezahn J, Weinhold R, Gerbaulet C, *et al.*: **Conventional power plants.** 2016.
Publisher Full Text

40. **JRC Hydro-power plants database.** Contribute to energy-modelling-toolkit/hydro-power-database development by creating an account on GitHub, Energy Modelling Toolkit, 2019; (accessed April 18, 2019).
Reference Source

41. European Commission, Joint Research Centre: **ENSPRESO - SOLAR - PV and CSP.** European Commission, Joint Research Centre (JRC) [Dataset] PID: 2019.
Reference Source

42. Walch A, Castello R, Mohajeri N, *et al.*: **Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty.** *Appl Energy.* 2020; **262**: 114404.
Publisher Full Text

43. Bódis K, Kougias I, Jäger-Waldau A, *et al.*: **A high-resolution geospatial assessment of the rooftop solar photovoltaic potential in the European Union, Renew.** *Sustain Energy Rev.* 2019; **114**: 109309.
Publisher Full Text

44. **OpenStreetMap contributors.** Planet dump retrieved from
Reference Source

45. European Commission, Joint Research Centre: **ENSPRESO - WIND - ONSHORE and OFFSHORE.** European Commission, Joint Research Centre (JRC) [Dataset] PID. 2019.
Reference Source

46. Bundesamt für Energie: **Konzept Windenergie - Grundlagenkarte des Bundes betreffend die hauptsächlichen Windpotenzialgebiete**. 2017. (accessed March 1, 2021).

47. Norwegian water resources and energy directorate (NVE): **Offshore wind power in Norway Strategic environmental assessment– English summary**. 2013. (accessed March 1, 2021).
Reference Source

48. Kies A, Chattopadhyay K, Lorenz E, *et al.*: **Simulation of renewable feed-in for power system.** Institute of Physics, University of Oldenburg, 2016.
Reference Source

49. ENTSO-E: **Power Statistics.** (n.d.). (accessed March 2, 2021).
Reference Source

50. Ruhnau O, Hirth L, Praktiknjo A: **Time series of heat demand and heat pump efficiency for energy system modeling.** *Sci Data.* 2019; **6**(1): 189.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Duque JC, Anselin L, Rey SJ: **THE MAX-*P*-REGIONS PROBLEM*.** *J Reg Sci.* 2012; **52**(3): 397–419.
Publisher Full Text

52. Hülk L, Wienholt L, Cußmann I, *et al.*: **Allocation of annual electricity consumption and power generation capacities across multiple voltage levels in a high spatial resolution.** *Int J Sustain Energy Plan Manag.* 2017; **13**: 79–92.
Publisher Full Text

53. Robinius M, Ter Stein F, Schwane A: **A Top-Down Spatially Resolved Electrical Load Model.** *Energies.* 2017; **10**(3): 361.
Publisher Full Text

54. Fleischer C: **ENSYSTRA/EU-SES: v1.3**. *Zenodo*, 2022.
http://www.doi.org/10.5281/zenodo.5834185

55. Scarlat N, Dallemand JF, Fahl F: **Biogas: Developments and perspectives in Europe.** *Renew Energy.* 2018; **129**: 457–472.
Publisher Full Text

56. Hundleby G, Freeman K: **Unleashing Europe's offshore wind potential - A new resource assessment**. 2017.
Reference Source

57. Fleischer C: **A data processing approach with built-in spatial resolution reduction methods to construct energy system models**. 2022.
http://www.doi.org/10.17605/OSF.IO/JHMXN

58. Moles C, Sigfusson B, Spisto A, *et al.*: **Energy Technology Reference Indicator (ETRI) projections for 2010-2050**. 2014.
Reference Source

59. Jülch V: **Comparison of electricity storage options using levelized cost of storage (LCOS) method.** *Appl Energy.* 2016; **183**: 1594–1606.
Publisher Full Text

60. **COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS**. The European Green Deal. 2019.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 03 March 2022

https://doi.org/10.21956/openreseurope.15497.r28570

© **2022 Tröndle T.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Tim Tröndle**
Institute for Environmental Decisions, ETH Zürich, Zürich, Switzerland

I approve this revised version of the article.

I have one remaining comment. I still think the introduction does not make the literature gap clear to the reader. It mentions existing tools but does not explain why they are insufficient or how this new approach expands their possibilities.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Energy systems, open data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 March 2022

https://doi.org/10.21956/openreseurope.15497.r28571

© **2022 Usher W.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **William Usher**
Department of Energy Technology, KTH Royal Institute of Technology, Stockholm, Sweden

In my opinion, the revised article is ready for acceptance.

I have spotted one small error - there is a missing reference on page 6 beginning the sentence "[]

provides the historical daily inflow in 30 European countries..."

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: energy system modelling; Python development; data processing workflows;

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 11 August 2021

https://doi.org/10.21956/openreseurope.14490.r27282

© 2021 Usher W. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

William Usher
Department of Energy Technology, KTH Royal Institute of Technology, Stockholm, Sweden

Overview
The paper presents a detailed description of the data processing required to conduct an analysis of European energy system. The data is aggregated in three different ways and the output from this process is then used to generate three alternative energy system models using the Calliope framework. The results differ between the models, although the explanation for these differences is insufficiently detailed.

I believe this paper presents a useful addition to the research literature. In particular, the description of the data processing workflow is very detailed, providing clear guidance to others who wish to replicate or build upon this work. There is research-value in this data processing, but I would expect more critical reflection on the choice of data sources and implications of these choices. In addition, there is a good opportunity to investigate the modifiable areal unit problem (MAUP) and ecological fallacy – classical challenges in the aggregation of spatial data.

At present, the data processing parts stops abruptly with no reflection on the process.

The results and discussion almost exclusively focus on the results from the three models, but it is unclear in this section whether differences in the model results are derived from the choice of modelling framework, or from the aggregation of the data, or the interaction of the two (reviewer 1 highlights an example of this). The results seem rather weak or vague: “the transmission capacity expansion limits *may have* resulted in...” and “demonstrate the importance of spatial context in energy system modelling” and should be strengthened.

Recommendations

I suggest that the authors refocus the paper on the data processing workflow, using an abbreviated version of the existing modelling results section as a case study to demonstrate the implications of different aggregation approaches.

The introduction and results & discussion sections could be reformulated to highlight:

- The research gaps that the paper fills:
  - the need for high spatial and temporal analysis in energy.
  - there's a lot of spatial data out there, but few papers which demonstrate how to bring it together.
  - how to avoid pitfalls when doing this.
- Whether existing data sets are sufficient, and where improvements are needed.
- What extra processing is required and implications of the assumptions and techniques used (such as for the derived variables).

The modelling case study should then reinforce some of these important points, for example demonstrating the response of the model to different aggregation methods.

Other comments

Grammatical niggles: "Since the early 2000s, energy system models *have become* more openly available." "There are a limited *number* of available open-source modelling tools…"

The fact that the paper is unpinned with a Python package (EU-SES) to reproduce the results is not clearly highlighted separately from the energy system modelling. Clarifying the generalisability of the package would be useful e.g. how easy is it to use it with other energy system models? The following lines from the readme could be sufficient for this paper?

```
import euses
# First specify countries and year of interest
# Use nomenclature of countries given in euses/parameters.py
year = 2010
countries = ['Austria','Switzerland']

# The build_dataset function builds the xarray areas dataset
example = euses.build_dataset(countries,year)
example.create_regions('poli_regions')
```

Derived variables: Discuss implications of electricity load profiles using population shares. E.g. errors introduced in low populated areas with high industrial demand?

Could you show an example CO2 pathway of the cumulative cap of 20% of 1990 levels? How does this relate to policy?

Results and discussions: "This is despite the fact that Germany is represented…" – use of "despite" seems incorrect, as I would expect a more aggregate model to omit detail on solar.

As highlighted by reviewer 1, the rooftop solar result comes out of nowhere. If retaining this in the case study, it should follow from a research question.

Figure 3d – clarify if % of available area used is absolute or relative to the 50% cap you impose – this is confusing to a reader.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: energy system modelling; Python development; data processing workflows;

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Jan 2022

Christian Fleischer

Author's response to the review by William Usher **REPLY TO REVIEWER 2 COMMENTS** – (Reviewer's comments in italics)

I suggest that the authors refocus the paper on the data processing workflow, using an abbreviated version of the existing modelling results section as a case study to demonstrate the implications of different aggregation approaches. The introduction and results & discussion sections could be reformulated to highlight:

- *The research gaps that the paper fills:*
  - *the need for high spatial and temporal analysis in energy.*
  - *there's a lot of spatial data out there, but few papers which demonstrate how to bring it together.*
  - *how to avoid pitfalls when doing this.*
- *Whether existing data sets are sufficient, and where improvements are needed.*
- *What extra processing is required and implications of the assumptions and techniques*

used (such as for the derived variables).
The modelling case study should then reinforce some of these important points, for example demonstrating the response of the model to different aggregation methods.
Author reply: Following recommendations from reviewer 1 and reviewer 2 to focus the article on the data processing workflow, the article type of the article has been changed to a Methods article. The method article describes the data processing and provides three case studies to demonstrate various ways that the workflow can be applied to build energy system models. The data gaps and what improvements are needed are now discussed in the results and discussions section. While addressing some of these recommendations the data processing approach was improved. The first improvement was the use of temperature data at NUTS 2 level to calculate the coefficient of performance of air-sourced heat. The data processing approach now allows technology specific proxy variables to be used when calculating the weighting factors. The proxy variables are given Table 2.

Grammatical niggles: "Since the early 2000s, energy system models have become more openly available." "There are a limited number of available open-source modelling tools..." **Author reply:** This phrase has been reformulated to – "In the past, energy system models were primarily closed and proprietary. However, recently more open-source energy system modelling tools have been made available."

The fact that the paper is unpinned with a Python package (EU-SES) to reproduce the results is not clearly highlighted separately from the energy system modelling. Clarifying the generalisability of the package would be useful e.g. how easy is it to use it with other energy system models? The following lines from the readme could be sufficient for this paper? import euses # First specify countries and year of interest # Use nomenclature of countries given in euses/parameters.py year = 2010 countries = ['Austria','Switzerland'] # The build_dataset function builds the xarray areas dataset example = euses.build_dataset(countries,year) example.create_regions('poli_regions')"
Author reply: A jupyter notebook, called *Three case studies*, was created to allow for the replication of the case studies and the results presented. The following sentence has been included in the "Power and heat optimisation model development" section to elaborate on the how easy is it to use it with other energy system models - The statement reads as, "The current version of the EU-SES tool can only automate the construction of an energy system model using the calliope framework. However, as the datasets are separated from models, the datasets can be used as input data in other modelling frameworks such as PyPSA." Also, to reduce the amount of time needed to solve the case study models the models were simplified in two ways. The first simplification is the removal of the battery technology in the model. The second simplification is the setting of the transmission capacity to a fixed value of 2GW and cannot be reduced or increased. The reference year used to build the Areas dataset for the case studies has been changed from 2010 to 2011.

Derived variables: Discuss implications of electricity load profiles using population shares. E.g. errors introduced in low populated areas with high industrial demand?
Author reply: In the results and discussion section a few sentences are dedicated to address the limitations to using population data to disaggregate power profiles. Additionally, In the derived variable subsection, some references to studies that use population, GDP or a combination of population and GDP to disaggregate power profiles.

Could you show an example CO2 pathway of the cumulative cap of 20% of 1990 levels? How does this relate to policy?
Author reply: To address this comment the following paragraph has been added to the article: "The cumulative CO_2 emission constraint ensures that the models have high solar and wind penetration levels. This emission constraint aligns with the roadmap presented in 2011 by the European Commission that aims to reduce 80 % of the EU CO_2 emission by 2050. In 2019 the EU commission revised the CO_2 emission target for 2050 to a net-zero emission target[10]. Therefore the 80 % reduction target could represent a snapshot along the net-zero pathway.

Results and discussions: "This is despite the fact that Germany is represented..." – use of "despite" seems incorrect, as I would expect a more aggregate model to omit detail on solar. **Author reply:** With the change in article type the Results and discussion section has been edited to contain case studies that use the EU-SES modelling tool.

As highlighted by reviewer 1, the rooftop solar result comes out of nowhere. If retaining this in the case study, it should follow from a research question.
Author reply: The models with 50% rooftop solar PV cap has been omitted as the article now focuses on describing the method and presenting some case studies.

Figure 3d – clarify if % of available area used is absolute or relative to the 50% cap you impose – this is confusing to a reader.
Author reply: The models with 50% rooftop solar PV cap were omitted therefore the available area is relative to the absolute available PV area.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 May 2021

https://doi.org/10.21956/openreseurope.14490.r26739

© 2021 Tröndle T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Tim Tröndle
Institute for Environmental Decisions, ETH Zürich, Zürich, Switzerland

This work describes a data-processing pipeline which can be used to generate input data for models of the European energy system. The pipeline is used to build three models with different spatial scope and spatial resolution. All three models are then run twice: once with and once without the availability of rooftop photovoltaics. The main finding is that model results show a stronger change in one of the models.

I believe the data-processing pipeline has the potential to be a valuable contribution to the

scientific community but this article requires revisions. My main concerns are (1) mixing the data-processing pipeline with the actual energy system model, (2) the unclear research gap, (3) the conclusions, and (4) the relevance of the research question. I describe these points in more detail below.

To me, the strongest part of this article is the data-processing pipeline. I suggest the author considers publishing it as a Data Note on this platform rather than a Research Article. Most of my points below regard the research part of this article and thus are invalid when published as a Data Note without the analysis.

1. The title, introduction, and large parts of the method section suggest that this article focuses on the data-processing step of energy system modelling: the pre-processing of raw input data that is necessary to build models. To test the impact of spatial resolution and scope, the author applies output of the data-processing step to energy system models and discusses model results. Mixing data-processing with the actual model obscures causal mechanisms and makes it difficult to attribute findings to their source. In fact, some of the findings are artefacts of the energy system models rather than the data-processing pipeline. For example, the author finds more transmission capacities in optimised systems with higher spatial resolution. This is due to a model decision of limiting the maximum installable transmission capacity independently of spatial resolution and it is therefore a model artefact rather than a consequence of the choice of spatial resolution of the data-processing pipeline.

   I suggest to make the impact of modelling choices on findings clearer in this article. In fact, I believe this article was stronger if it did not apply any energy system modelling but directly compared outputs of the data-processing pipeline.

2. The introduction mentions two existing data-processing approaches with the same purpose but does not mention any gaps in the existing approaches. I suggest to make it clear to the reader what the added benefit of this new approach is.

3. Modelling German city-states in isolation shows the largest sensitivity towards the lack of rooftop photovoltaics. This result is not surprising as the larger part of solar power stems from rooftops in a city-state like Berlin. I assume the model with lower spatial resolution leads to the exact same effect, but that effect is less visible as inputs and outputs are aggregated over larger areas. Is my assumption correct? If yes, what meaning do the findings have for modellers?

   I would like to verify this assumption myself, but I can not because model results are not published. Results from the models GER MAX-P and GER NUTS0 should be published in the same way results from GER NUTS1 are.

4. While not a criteria on this platform, I strongly recommend to improve the relevance of the research question. The study investigates the impact of a complete loss of rooftop solar potential in Germany using energy system models of different spatial scope and resolution. The relevance of this question and the choice of the analysis method is not clear to me. Why is this question relevant? What are the implications for further research?

Additional minor comments:

- Figure 2 seems to be incomplete, as the text mentions other storage options than the one

shown. I suggest to add these options to the Figure or mention their omission in the caption.

- The JRC hydro database has been updated and improved several times since the reported access date. I suggest to use the most recent version in the pipeline as it resolves some of the problems mentioned.

- The study 16 has four spatial scales: continental, national, regional, and municipal.

- Zenodo can host datasets that consist of more than one file. The section “Dataset framework” seems to suggest the opposite.

- Some references in the section “Base variables“ point to the wrong source. For example, the OPSD platform is referenced with number 34, but the reference leads to an unrelated JRC publication.

- Some references need revision:
  - 6. Should cite the article rather than the pre-print.
  - 11. Should cite the article rather than the pre-print.
  - 16. Should cite the article rather than the dataset that is not used in this study.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Energy systems, open data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Jan 2022

Christian Fleischer

Author's response to the review by Tim Tröndle **REPLY TO REVIEWER 1 COMMENTS** – (Reviewer's comments in italics)

I believe the data-processing pipeline has the potential to be a valuable contribution to the scientific community but this article requires revisions. My main concerns are (1) mixing the data-processing pipeline with the actual energy system model, (2) the unclear research gap, (3) the conclusions, and (4) the relevance of the research question. I describe these points in more detail below. To me, the strongest part of this article is the data-processing pipeline. I suggest the author considers publishing it as a Data Note on this platform rather than a Research Article. Most of my points below regard the research part of this article and thus are invalid when published as a Data Note without the analysis.
Author reply: As suggested the article type has been changed to focus the data processing pipeline. Although the reviewer suggested to change the article type to a Data Note type, following consultation with Open Research Europe editorial a Methods article type was found to best fit the content of the article. With the change in article type, the points by the reviewer that address the analysis are not addressed.

Figure 2 seems to be incomplete, as the text mentions other storage options than the one shown. I suggest to add these options to the Figure or mention their omission in the caption. **Author reply:** The hydrogen storage option has been added to Figure 2. Note that the battery has been removed as a storage option in the models for simplification purposes and is therefore not included in Figure 2.

The JRC hydro database has been updated and improved several times since the reported access date. I suggest to use the most recent version in the pipeline as it resolves some of the problems mentioned.
Author reply: The version of the JRC hydro database has been updated to V10 (most recent version on January 6th 2021). V10 of the database still has the issue of not providing storage capacity of the hydropower plants with storage capabilities.

The study 16 has four spatial scales: continental, national, regional, and municipal.
Author reply: The municipal spatial scale has been added to the article.

Zenodo can host datasets that consist of more than one file. The section "Dataset framework" seems to suggest the opposite.
Author reply: To correct the suggestion made in this section the phrase has been reformulated to state, "The datasets created can be stored and shared using a single file in online archives such as Zenodo."

Some references in the section "Base variables" point to the wrong source. For example, the OPSD platform is referenced with number 34, but the reference leads to an unrelated JRC publication.
Author reply: The reference for conventional power plants dataset hosted on the OPSD platform has been corrected with the correct reference number 37. Other misplaced references were not found in the Base Variable sub-section.

Some references need revision: - 6. Should cite the article rather than the pre-print. - 11. Should cite the article rather than the pre-print. - 16. Should cite the article rather than the dataset that is not used in this study."
Author reply: Suggested corrections have been made.

Competing Interests: No competing interests were disclosed.