Harmonized and Open Energy Dataset for Modeling a Highly Renewable Brazilian Power System

Deng, Ying[1,*], Cao, Karl-Kiên[1], Hu, Wenxuan[1], Stegen, Ronald[1], von Krbek, Kai[1], Soria, Rafael[2], Rochedo, Pedro Rua Rodriguez[3], and Jochem, Patrick[1]

[1]German Aerospace Center (DLR), Institute of Networked Energy Systems, Curiestr. 4, 70563 Stuttgart, Germany
[2]Department of Mechanical Engineering, Universidad San Francisco de Quito, Diego de Robles y Vía Interoceánica, Campus Cumbayá, 170901 Quito, Ecuador
[3]Energy Planning Program, Graduate School of Engineering (COPPE), Universidade Federal do Rio de Janeiro, Centro de Tecnologia, Bloco C, Sala 211, Cidade Universitaria, Ilha do Fundão, 21941-972, Rio de Janeiro, Brazil
[*]corresponding author(s): Deng, Ying (ying.deng@dlr.de)

ABSTRACT

Improvements in modeling energy systems of populous emerging economies are highly decisive for a successful global energy transition. The models used – increasingly published open source – still suffer from the lack of appropriate open data. As an illustrative example, we take the Brazilian energy system, which has great potential for renewable energy resources but still relies heavily on fossil fuels. We provide a comprehensive open dataset for scenario analyses, which can be directly used with the popular open energy system model PyPSA and other modeling frameworks. It includes three categories: 1) time series data of variable renewable potentials, electricity load profiles, inflows for the hydropower plants, and cross-border electricity exchanges, 2) geo-referenced data for 27 defined regions, and 3) tabular data, which contains power plant data with installed and planned generation capacities, aggregated grid network topology, biomass thermal plant potential, as well as scenarios of energy demand. This data fosters global or country-specific energy system studies based on open data relevant to decarbonizing Brazil's energy system.

Background & Summary

The decarbonization of energy systems in developing countries, especially in the most populous ones, becomes a determinant factor for a global "well below 2°C" target[1]. Achieving climate neutrality requires complete or nearly complete decarbonization of the electricity system. This is attainable today through a number of technologies that provide low-carbon or even carbon-free electricity — renewable energy, nuclear power, and fossil-fueled electricity with carbon capture and storage. Low social acceptance and low economic viability make the latter two technologies more challenging to deploy on a large scale, and their timely installation questionable. However, the generation profile and production costs of variable renewable energy sources (vRES) vary with the weather, i.e., the spatial location and the availability of wind resources and solar radiation. As a consequence, the decision problems in the operation and planning of reliable, stable, and carbon-neutral power systems rely on large-scale models and datasets.

Open science promotes the use of open models for supporting the transition to carbon-neutral energy systems. Typically, such open models are populated with datasets specific to the power system. However, energy data can come from different sources, and the accessibility and licensing conditions of energy data affect the degree of openness of the modeling workflows[2]. For this reason, the open data can help drive and support the efforts of improving transparency and productivity[3]. In developed countries, especially in Europe, there are various open energy system models that are available as open source[4]. At the same time, there are several platforms, for instance, the Open Energy Platform [i] and Open Power System Data platform[5] which coordinate various open datasets related to climate, demand profiles, transmission grids, scenarios, etc. for modeling the European power system. In contrast, energy system models for developing countries use datasets that are opaque and, in most cases, inaccessible. This makes it difficult for global energy models to adequately represent developing countries. Language barriers may further hinder researchers who do not belong to the same language region from utilizing available energy data. As one of the five most populous countries, Brazil is a developing country with significant wind resources and solar radiation potential, albeit in the early stages of deployment. Brazil's energy system is facing a strategic transition and its capacity expansion is constrained by the rainforest. All this makes it valuable to understand the Brazilian energy system itself in detail

[i]https://openenergy-platform.org/

and its potential contribution to the global energy transition.

In this context, our contribution is to make the existing energy data of Brazil better applicable for energy systems modeling. By providing the first publicly available, spatially explicit, harmonized, and English version of Brazil's energy data, we enable researchers to replicate the Brazilian energy system and/or to improve the integration into global energy models starting from a common basis.

The assembled dataset comprises the following subcategories as detailed in the Methods section: i) map of the defined region, ii) aggregated grid network topology, iii) vRES potentials – profile and installable generation capacity, iv) geographically installable capacity of biomass thermal plants, v) hydropower plants inflow, vi) existing and planned power generators with their capacity, vii) electricity load profile, viii) scenarios of sectoral energy demand and ix) cross-border electricity exchanges. This dataset is resolved geographically by Brazilian federal states and time series data are resolved by hours, spanning the period 2012-2020.

In this way, the presented dataset provides the basic information and foundation for the operational and expansion planning studies necessary to explore a highly decarbonized energy future for Brazil. For example, the dataset was used in the PyPSA-Brazil model[6] to assess the impact of transmission grid expansion in the Brazilian power system. The dataset published in this paper has been updated and includes more years of data than the version[6] used.

Methods

The purpose of this work is to create consolidated open energy data for Brazil based on open and accessible original datasets. Table 1 summarizes the sources and licenses of the raw data used for each subcategory of the dataset in this paper.

The following subsections elaborate on 1) the knowledge of energy data in the Brazilian context, 2) how each dataset is obtained from its original source, and 3) the assumptions made in the processing and construction of the datasets.

Defined region

Brazil is a country with 5 macroeconomic regions, 4 electric regions, 27 federal levels (26 states and one federal district – Brasília), and 5572 municipalities.

The spatial resolution of the dataset is at ISO 3166-2 level [ii] and comprises 27 defined regions, i.e., federal level, illustrated in figure 1.

Data collection

Even though there are several map sources, the original dataset used is from the Brazilian Institute of Geography and Statistics (Portuguese: Instituto Brasileiro de Geografia e Estatística; IBGE)[7]. This choice is not only motivated by the licensing, but also because IBGE is the official map source for Brazil and is considered to be the most credible source for the country's borders and topography. The Coordinate Reference System (CRS) of the shapefile is SIRGAS 2000 (commonly known as EPSG:4674).

Data processing

These attributes in the original dataset[7] are converted to English and the CRS is re-projected to EPSG:4087. Only the federation state and the geometric information of the polygon are retained. In addition, representative coordinates (x, y) of the federal states are added and are considered as the centroid of the state polygon.

Aggregated grid network topology

The power grid connects all power generators and loads. In Brazil, the electricity grid is known as the National Interconnected Network (Portuguese: Sistema Interligado Nacional, SIN) and is managed by the National Electricity System Operator (Portuguese: Operador Nacional do Sistema Elétrico, ONS). ONS divides Brazil into four electric regions, each of which includes several federal states, as shown in table 2. SIN has a total length of 167,000 km and connects almost the entire country (96.6% of the national territory), except for some isolated places in the northern region. Over the next few decades, 434 lines with a total length of 32,000 km are planned to be built[8].

Data collection

The complete grid topology of Brazil is taken from the dataset published by Energy Research Company (EPE), called EPE Webmap[8]. The original datasets are available in the form of shapefiles, with transmission line data as the line layer and substation and generator data as the point layer. All lines, substations, and power plants have individual shapefiles, which are classified by their operational status – existing or planned – and, for power plants in particular, by their plant type. The CRS of the shapefile is SIRGAS 2000. The attributes are specified in Portuguese and include name, plant operator, voltage level, year of operation, and line length, among others. Substations, transmission lines, and different types of power plants shapefiles are used to derive the network topology. Table 3 lists the number of records in the original datasets used.

[ii] ISO 3166 refers to the code indicating the name of the country and its subdivisions https://www.iso.org/iso-3166-country-codes.html

Data processing

Each federal state is modeled as a node located in its geometric center and connected by transmission lines currently in place and in the national ten-year plan[9]. We assume that existing and planned transmission lines are operating regardless of the scenario year, so we add up the transmission capacity and ignore the reference year. Information on the connection of the lines to substations or power plants is not provided in the original data, however, this is necessary to construct the grid topology. For this purpose, heuristics are used to connect the starting and ending points of transmission lines to nearby substations or power plants. The processing is divided into three parts: four steps of pre-processing, mapping, aggregating and representing, as displayed in figure 2.

Before the mapping action, there are four pre-processing steps to make the "spatial join at the closest distance" algorithm effective.

1. The federal states to which the substations and power plants belong are added to the attribute table according to their geographical locations.

2. Information on existing foreign substations connected to the SIN is added manually based on[9]. This is due to the fact that the transmission lines indicated in the original line layer contain international connections, while information about substations located outside Brazil is not specified. Added attributes include the name of the substation, the operator, the voltage, and the geometry. In addition, a new attribute, namely state, is added to identify the country to which it belongs using the ISO 3166-1 alpha-3 code. The state of the substation abroad is three characters, whereas in Brazil it is two characters. The geometry added manually is the longitude and latitude where the substation is located. An exception is the SE Macagua substation, located in Venezuela. Its actual location is (8.304, -62.668). However, it is designated as (4.530, -61.138). This is because, in the original data, the transmission line to the Boa Vista substation ends here. Additionally, the heuristic algorithm is based on the nearest distance criterion.

3. LineString in the transmission line layer has to be further processed by converting MultiLineString to LineString and closed LineString to open LineString.

4. The shapefiles are reprojected to EPSG:4087 so that the distance-based calculations are robust.

After pre-processing, the "sjoin_nearest" function in the Geopandas package in Python is used to map between the geometry of the start and end points of the line layer and the geometry of the substations and power plant points. The maximum distance to query the nearest geometry starts from an initial distance of 1 km and is incremented by 1 km at the next query. Table 4 reveals the statistics of the mapping results, where sub_0 represents the start point and sub_1 represents the endpoint. More than 90% of the mappings (96.1% of the starting points and 94.4% of the ending points) are found to be within 1 km. The greatest deviation in mapping is caused by the line LT 230 Itapaci – Mineradora Maracá, with a length of 85 km, especially the mapping of its endpoints, since the nearest points of the line's start and end points are the Itapaci substation.

The final step is to aggregate these lines to represent the network topology between each federal state. Depending on the federal state information, only trans-state transmission is selected, which assumes that potential grid bottlenecks are not considered inside the defined region – "copper plates" assumption[10]. The original dataset does not have information on whether the lines are alternating current (AC) or high-voltage direct current (HVDC) lines. There are several duplicate records for HVDC lines, such as Porto Velho - Araraquara, and Xingu - Estreito, hence these records are removed. The transfer capacity of the HVDC lines is supplemented manually with information from various sources, as specified in table 5.

To calculate the transfer capacity of AC lines, the number of circuits in each transmission line is added. "C1" and "C2" in the line names represent the first circuit and the second circuit, marking each line of the parallel circuit, while "CD" indicates a double circuit[11]. Therefore, each line defaults to a single circuit, while lines with a "CD" tag in the line name are set to a double circuit. However, the original dataset had no information on the physical characteristics of the lines, such as the conductor resistance, inductance, and capacitance of each transmission line. Therefore, we assume that each line is 4 bundles of conductors. The remaining transmission lines of different voltage levels are unified as parallel lines of 380 kV, thus forming an equivalent transmission network. This enables the transmission capacities starting in the same federal state and ending in another identical federal state to be added up. To calculate the transmission capacity of the equivalent transmission system, the lines are assumed to be three-phase overhead lines and of type 490-AL1/64-ST1A[12].

The transfer capacity (apparent power) is calculated:

$$S = \sqrt{3}UI, \tag{1}$$

The transmission losses

$$f = 1 - \frac{3\mathrm{R}_L' I^2 n}{S}, \tag{2}$$

where:

$$
\begin{aligned}
S &= \text{apparent power, MW} \\
U &= \text{voltage level, kV} \\
I &= \text{nominal current of the wire, kA} \\
R_L' &= \text{DC resistance rating of the conductor at operating temperature for the wire, } \Omega/\text{km} \\
n &= \text{number of bundle conductor, n=4} \\
f &= \text{the transfer efficiency considered as 1 minus the effective loss of each line.}
\end{aligned}
$$

In aggregation, transmission capacity is accumulated and efficiency and line length are averaged out. In figure 3, the results are illustrated.

Power plants infrastructure

Generators are an integral part of the energy industry, responsible for producing electricity and injecting it into the grid – transmission and distribution – so that it reaches consumers.

Data collection

There are several official generator databases in Brazil, for example, ANEEL-SIGA[13] published by National Electric Energy Agency (ANEEL), EPE Webmap[8], ONS Historical Database [iii]. ANEEL-SIGA is the Generation Information System and contains information on power plants from the granting phase to the decommissioning phase. EPE Webmap refers to the Geographic Information System of the Brazilian Energy Planning Studies. It is a geo-referenced database containing official information for medium and long-term energy planning in Brazil. The power plants in the ONS Historical Database mainly refer to those which are operated by ONS and are part of its SIN. In general, when a power plant is in operation, it implies that it is connected to the SIN. There are some exceptions, such as isolated systems, supplied by local generations and not connected to the SIN. The types of plants considered in the three datasets are different, as shown in table 6. To enable comparison between datasets, the installed capacities are summed to the plant types defined by ONS, and the results are shown in table 7.

Table 7 shows that the number of plants is evidently different, while the total installed capacity of each type of plant is similar, except for the thermal plants. In particular, the total installed capacity of ONS and EPE Webmap is similar, while the ANEEL-SIGA dataset includes more thermal capacity, but relatively less wind and solar capacity. Different numbers of power plant units but similar total installed capacity may result from the following reasons:

1. EPE Webmap covers mainly centralized generation, whose operating mechanisms are self-generation and public utilities. In addition to the plants in the EPE Webmap, the ANEEL-SIGA database includes distributed generation under the net metering scheme and small-scale backup generators. ONS Historical Database contains the plants dispatched in SIN.

2. The dataset updates between ONS Historical Database, EPE Webmap, and the ANEEL-SIGA database are not synchronized. ONS publishes information on operating power plants on an annual basis, and the latest update of the EPE Webmap is September 2020. On the contrary, the ANEEL-SIGA database is constantly updated with the granting of power plants. However, the historical versions of ANEEL-SIGA are not accessible.

3. The different definitions of plant units. ANEEL documents each plant unit based on when they received their grant, while ONS defines projects based on their operating units.

Unlike the other two datasets, the ONS Historical Database does not provide geographic coordinates, but rather the electric region in which the plant unit is located. Using this information, the installed capacity from different sources is aggregated to the electric region level and further compared for differences. Our results are presented in table 8. Comparing table 8 with table 7, the differences in the total installed capacity of each plant are further amplified at the regional level.

For the reasons mentioned above, unifying these three datasets into a single one is challenging. In addition, the names of the power plant units in the three databases are not aligned. In ANEEL-SIGA and EPE Webmap, power plant units can be uniquely labeled according to a single code for the generation unit – the CEG (company identifier). Table 9 explains its format. However, the CEGs of these two datasets do not match exactly. Where they do match, the installed capacity or other information, such as the basin of the hydropower plant, may differ.

[iii] The CCEE database published by Electric Power Commercialization Chamber (CCEE) is not studied because it is theoretically compatible with the post-2014 ONS Historical Database[14]

In total, there are 10,541 power plant units with 21 attributes in ANEEL-SIGA. From the database, these attributes include the name of the power plant, the plant ID [iv], federal state to which it belongs [v], city it belongs to, plant type, primary energy source, fuel type [vi], installed capacity, geographic coordinates of each generator, production capacity, primary fuel type, time in operation, phase-out time. The CRS used for the ANEEL-SIGA dataset is SIRGAS 2000, with coordinates expressed in degrees minutes seconds (DMS).

Data processing

Even though the ANEEL-SIGA data can be displayed online through PowerBI, it only provides a download link and there are slight inconsistencies between the downloaded files (in Excel format), e.g., plant coordinates, plant names, etc. Therefore, the dataset provided in this paper is based on the version downloaded by the authors on June 9, 2021.

Although there are differences between versions, the coordinates of the power plants in the dataset are not complete or the coordinates are not entirely within the range of Brazil. The coordinates are first converted to decimal form so that they can be easily re-projected. Since the record of city names is complete, the missing coordinates of the plant are assigned with the location of the city. An individual power plant unit located in more than one city can have multiple values in the 'city' property. For those plant units, only the first value of the city name is considered. In total, there are 847 records with missing coordinates or coordinates outside Brazil. Once the coordinates have all been replenished, information on the federal states is updated with the coordinates.

The installed capacity of each power plant unit determines its size. In the original dataset, the capacity is given in kilowatts and is provided separately for the purposes of granting, regulation and inspection. The granted capacity is the capacity considered in the act of granting, the regulated capacity corresponds to the capacity considered from the commercial operation of the first generating unit, and the actual guaranteed power is the average production in reality. Since not every power plant unit has information on the regulation capacity, the granted capacity is considered to be representative of the installed capacity. In addition, the units of installed capacity are converted to megawatts.

The information on the types of power plants in the original dataset is divided into eight types, summarized in table 10. There is a wave power plant (CGU), Porto do Pecém, installed in the state of Ceará, with a power of 0.05 MW, which is rebranded as small hydropower in this paper. Depending on the properties of the fuel source, thermal power plants are subdivided into oil-fired, natural gas-fired, coal-fired, and biomass-fired plants. In total, therefore, there are ten generic types of plants. Figure 4 illustrates the results of power plant distribution.

Most records have incomplete dates of commissioning and decommissioning. According to[15], the missing date information indicates that the plants are active. For those records showing the same commissioning date and decommissioning date, we set the decommissioning date to be missing. Finally, the reference year is added.

In the post-processing, the installed capacity of the power plants is aggregated to each federal state for each given reference year according to the type and operational status of the power plant. The installed capacity of power plants is aggregated, whether they are in public service, self-generating, or independent production. For the installed capacity of the reference year, it is assumed that the operational status of the plants should be operation – and that the commission time should either precede the base year or missing. Records in the original dataset that are in "construction" or "construction not started" status are renamed to "planning" status. Finally, the values are accumulated according to the federal state, plant type, and operational status. The installed capacity is different according to the reference year provided, while the planned capacity is the same for all reference years. This is because 68.8% of the records do not have a decommissioning date. Therefore, we tend to ignore this information.

Installable capacity for biomass thermal plants

Biomass can be burnt directly for heating or power generation, or it can be converted into oil or natural gas substitutes. Over the last 15 years, the generation of electricity from biomass thermal plants in Brazil has been increasing, from 6 GW to 14 GW, accounting for 13% of the capacity matrix of electricity for 2020. Sugarcane bagasse is the main source of biomass.

Data collection

No studies for Brazil investigating the energy production potential of biomass thermal plants are known to the authors, but[16] address the geographically installable capacity. In that paper, the authors estimate the technical, environmental sustainability, and economic feasibility of agricultural and agro-industrial residues. The economically viable potential refers to the power generation potential of biomass residues distributed within 50 km of the biomass thermal plant. According to their assessment, the total economic potential in Brazil is 39 $\mathrm{MWh/yr}$. With the authors' permission, they have provided us with the results of their paper in terms of economic potential in $\mathrm{MWh/yr}$, which are spatially resolved at the municipal level.

[iv] refers to CEG

[v] each record can be a single power plant or a power plant unit consisting of multiple power plants, for example, a wind farm operating multiple wind turbines. The federal state is determined by the location of the power plant units provided.

[vi] biomass, wind, fossil, hydroelectric, nuclear, wave

Data processing

The primary energy source used in today's biomass thermal plants is sugarcane bagasse [vii], and the share of residue is small to negligible. Therefore, only the economic potential of biomass from residues is considered as additional installable capacity beyond the already existing and planned installations.

As a first step, we convert the potential production into the additionally installable capacity by assuming an annual availability factor of biomass at 0.6[17]. Then the values are aggregated at the federal state level.

The biomass thermal plants included in the study[17] are from centralized plants published by ANEEL which are no longer accessible. We assume that the geographic distribution of the biomass thermal plants they considered is similar to that covered in the subsection Power plants infrastructure. Although the number of hours in a year depends on whether there is a leap year or not, we have assumed in our calculations that the number of hours per year has the same value, i.e., 8760 h. Therefore, we obtain the geographically installable capacity for each state:

$$C_i = \sum_i \left(\frac{PR}{f \cdot 8760h} + CI_i + CP_i \right), \tag{3}$$

where:

C = geographically installable capacity, MW

i = the federal state

PR = the residual potentials at municipality level, MWh

f = annual availability factor

CI = installed capacity of biomass thermal plants, MW

CP = planning capacity of biomass thermal plants, MW.

Since the installed capacity differs for each reference year, the geographically available installed capacity varies accordingly. Therefore, we provide data for each reference year.

Electricity load profiles

Future energy systems are likely to shift to electricity from renewable sources as the major source of energy. Therefore, the temporal distribution of energy consumption becomes increasingly relevant to the design of future energy systems as the share of vRES increases and consumption patterns change. At the same time, the spatial distribution of energy consumption becomes increasingly important as renewable energy generation and its consumption are becoming regionally asynchronous.

Data collection

ONS publishes hourly load profiles for the four electric regions it operates in SIN[18], while EPE provides annual sectoral electricity consumption or consumers for each federal state[19]. Table 2 indicates each electric region and the federal states it contains. The electricity consumption published by EPE is obtained from the distributor's billing system, while the ONS dataset is the measured dataset[14]. The hourly profile of ONS is for 1999-2020, while the EPE dataset is for 2012-2020 [viii]. However, the value of total power consumption provided by ONS is greater than that of EPE. The reasons for this difference are the physical losses and the physical representation in the SIN system[14]. The differences between the ONS and EPE datasets are illustrated in table 11, where also the regional differences are depicted.

Data processing

Both the ONS dataset and the EPE dataset are used.

In the ONS dataset, there is one missing value in the hourly time series for each electric region for each year except 2019 and 2020, with the largest number of missing values in 2014 with 25 (no data available for 2014-02-01). For the missing values, the values from one week earlier are used. In addition, six values are negative in the time series for the northern region. These values are trimmed to zero as we regard this as a gross error.

We use the EPE dataset as an allocation factor to decompose the ONS load profiles to the Brazilian federal state level and to the end-use sector level. Therefore, there are two allocation factors – by sectoral annual consumption and by sectoral annual consumers. This means that we assume that the seasonal, intraweekly, and intraday variations remain consistent across states belonging to the same electric region, but differ in magnitude.

[vii] Sugarcane bagasse refers to the dry pulpy substance remaining after grinding sugarcane in order to extract their juice.

[viii] The authors retrieved the dataset in April 2021

In the EPE dataset, end-use sectors are classified into eight categories – residential, industrial, commercial, rural, public sector, public lighting, public services, and own use[ix]. To be consistent with sectoral electricity consumption in the subsection Scenarios of electricity demand, annual electricity demand is aggregated according to[14] and[1] using the four end-use sectors defined by[1]. Table 12 specifies the sectoral mapping between the EPE dataset and this paper. We should note that when aggregating the sectors defined by EPE, none of them are considered to be the transportation sector. This is because in EPE, public transport services are counted as part of the commercial sector, while the transport of goods for industrial purposes is considered to be part of the industrial sector. As the EPE allocation factors are only applicable for the period 2012-2020, the time horizon for electricity consumption by sector in the federal states provided in this study is also only applicable for the period 2012-2020. Figure 5 illustrates the results of the electricity demand, which considers the physical loss in the SIN system.

Scenarios of energy demand

Energy demand scenarios facilitate a strategic assessment of possible pathways for long-term planning and their respective internal consistency and associated uncertainties. Sector-specific modeling allows variations in demand from different resources and sectors to be estimated on a national scale. However, diverse models, methods, and assumptions lead to different scenarios and represent research positions – conservative or optimistic, dependent on fossil or renewable energy.

For those energy studies that cite annual energy demand projections as model inputs, the narrative of the demand scenarios used is not clearly explained, such as[20]. This makes it difficult to interpret their results, especially whether they are consistent with the Paris Agreement or not. A good understanding of sectoral energy demand published in reputable studies allows researchers in energy system modeling to emulate the demand-related parameterization and manage uncertainties.

Data collection

There are numerous scenarios for the future energy demand of Brazil. The most famous are published in three studies: i) World Energy Outlook (WEO), ii) EPE's Long-term National Energy Plan (Portuguese: O Plano Nacional de Energia, PNE), and iii) Exogenous energy demand studied by COPPE researchers.

The WEO scenario of the International Energy Agency (IEA) is considered to be the most authoritative source of insights into the world's energy demand. It updates its sector demand scenarios annually, region by region. The latest WEO study for 2021[1], regarded as "WEO2021" in this paper, includes the sectoral energy demand for Brazil in the extended data, in CSV format. The WEO2021 addresses the demand scenario to 2050 with a five-year time span providing reference data of historical demand for 2010, 2015, 2019, and 2020.

EPE's PNE is considered a fundamental instrument for Brazil to outline the government's strategy regarding the expansion of the energy sector in the coming decades. PNE 2030 is the first integrated planning study until 2030, while in December 2020, EPE released PNE 2050[21], extending the horizon to 2050. Our comparison relies on the PNE 2050 study, referred to in this paper as "PNE2050". However, the PNE2050 does not provide numerical data but rather presents it in the form of a table or charts for each end-use sector. We, therefore, have to extract these values manually and create a CSV file accordingly. PNE2050 provides projections of sectoral demand every ten years (i.e. 2030, 2040, and 2050), using historical data from 2015 as a reference.

COPPE is the most prestigious research institute in Brazil that studies energy planning in Brazil and the world. We refer to their scenario studies as "COPPE". Out of the 133 scenarios provided by COPPE, we selected three scenarios, as they are so far the latest and have distinct transition paths. COPPE scenarios have 5-year-time-steps, however, the data we received only contain the years 2030, 2040, and 2050. Sectoral demand for 2010 and 2015 are used as the starting point for the scenario assumptions. In the following, the three COPPE scenarios are shortly described.

To enhance the transparency of energy scenarios[22], this work creates a matrix of energy demand scenarios. This matrix, shown in table 13 provides a summary of the main criteria used by previous studies to model final energy consumption scenarios up to 2050 in Brazil, following the comparisons described in[23]. Trend scenarios are considered, which maintain a level of effort in climate action similar to current policies and NDCs, and ambitious mitigation scenarios aligned with the global goals until the end of the century on the stabilization of the average temperature increase of the planet relative to pre-industrial times by 2 °C and 1.5 °C. These scenarios show the role that electrification, in the different sectors, may play to achieve climate goals. The electrification of the transport sector in Brazil is not as achievable as in other regions due to the important role that traditional and advanced biofuels can play. This is especially evident in the lowBECCS scenario.

Data processing

We first normalize the units of demand values for the three studies to PJ because they are different in the raw data, i.e. PJ for the WEO2021, Mtoe[x] for PNE2050 and EJ for COPPE. After that, we give aliases in a format of `XXXX_YYYY` to represent the

[ix] These are the terms used in the original dataset. Own use refers to the self-consumption of the energy system.
[x] million tonnes of oil equivalent

studies and the corresponding scenarios. For example, the alias `COPPE_BAU` represents the Business as usual (BAU) scenario for the publication of the COPPE studies.

We align the end-use sectors and energy carriers in PNE2050 and COPPE with WEO2021 based on[1,21,24–26], as the different definitions prevent comparisons between them. Table 14 describes the correspondence. In fact, WEO2021 does not provide a value for the end-use sector named "Other". We assume that the value for the end-use sector "Other" is the difference between total final consumption (TFC) and sectoral demand:

$$Other_i = TFC_s - D_{s,i}, \tag{4}$$

where:

$Other_i$ = the energy demand for the energy carrier i, the end-use sector "Other",

TFC_s = total final consumption for end-use sector s,

s = end-use sector s, $s \in \{\text{Transport}, \text{Industry}, \text{Buildings}\}$,

i = the energy carrier i, $i \in \{\text{Total liquids}, \text{Total gases}, \text{Total solid fuels}, \text{Total}\}$,

$D_{s,i}$ = the energy demand for the end-use sector s and the energy carrier i.

PNE2050 data provides the most granular energy carriers, followed by WEO2021, while the COPPE scenarios divide the energy carriers into "electricity", "liquid", "gas", "solid" and "hydrogen". Table 14b list all energy carriers for PNE2050. The WEO2021 scenario dataset includes TFC, the total value of energy carriers by physical state, i.e., "total liquids", "total gases", "total solid fuels", as well as some of the more subdivided energy carriers. For instance, "total liquids" consists of "oil products", "liquid biofuels", and "hydrogen-based liquid fuels"[1]. However, "liquid biofuels" are not provided. Although an energy carrier "hydrogen" is provided in the COPPE scenarios, all scenarios have zero values. Therefore, we left the "hydrogen" out. Even when hydrogen as final energy is zero, there is an important hydrogen production as an intermediate energy carrier, which is input to produce other final energy forms. This intermediate production is not reported.

Figure 6a presents the total final energy consumption by the combined sector for "Transportation", "Industry", "Buildings", and "Others". The COPPE and EPE scenarios do not report the consumption of "Others". EPE's PNE2050 indicates that the three reported sectors account for more than 80% of final energy consumption and will continue to be just as important in the long term. "Others" basically considers energy consumption associated with Agriculture and Livestock. The average final energy consumption for the three most important sectors in 2015 was 7.9 PJ. This value rises to 10 PJ in 2030, and 12.4 PJ in 2050. In the long-term, there are important variations depending on the scenario (for more details of each scenario see table 13).

Figure 6b shows the total final energy consumption by integrated energy type for each scenario considered. In the long term, electrification is increasingly important in the three sectors with the highest consumption. Electricity consumption represented 19% in 2010, and the average between the scenarios indicates that it could reach 21% in 2030 and 28% in 2050. There are important differences in the role that electrification could play between scenarios, especially in the transport sector, where electrification decreases depending on the development that BECCS technologies may have in the long term. With a significant development of BECCS, total electricity consumption would represent approximately 1.8 PJ in 2050, while with a conservative development of BECCS, total electricity would represent approximately 4.4 PJ. For more details on the consumption of other solid, liquid, and gas energy, see table 13.

Inflow of hydropower plants

Hydropower is an essential sustainable energy source, especially in developing countries, such as Ecuador, China, and Brazil, where it accounts for the largest share of renewable energy sources and even of the total generation matrix. With the increasing penetration of vRES in the power system, the proper representation of hydropower in the power system analysis becomes crucial. The reason for this is that hydropower can be operated like storage, thereby increasing the flexibility of the power system. The potential production of hydropower is determined by the combination of available water flow and available head height at each location[27]. The power output is usually limited to the nameplate capacity of the hydroelectric plant at the maximum flow rate of the turbine.

The SIN is operated by ONS, which operates the hydraulic reservoir system. In total, ONS dispatches 163 plant units of different type, including 10 reservoirs, 92 run-of-river units, 60 hydropower plant units with reservoirs,s and 1 pumped hydropower plant. Hydropower plants with reservoirs in Brazil can provide about 210 TWh [xi] storage energy, of which about 69% is located in the SIN region of southeast/central Brazil, followed by the northeast region with about 18%, and the southern and northern SIN regions with 7% and 6%, respectively[28].

[xi] the used unit in the original dataset is MWmês. 1 MWmês = 720 MWh/month

Data collection

ONS publishes daily, weekly, and monthly resolved time series separately in relation to the inflow of the reservoirs – Affluent Natural Energy (Portuguese: Energia Natural Afluente, ENA) and Stored Energy (Portuguese: Energia Armazenada, EAR)[29]. These datasets are available at different levels of aggregation, i.e., by the reservoir, by subsystem, by basin, or by equivalent energy reservoir (Portuguese: Reservatório Equivalente de Energia, REE). These data are continuously updated.

ENA refers to the energy flowing to the hydropower system at aggregated levels. The EAR is a value that reflects the reservoir levels and how much energy they can still produce. The ENA and ERA datasets have been used in several studies, such as[30] and[31]. However, due to the lack of metadata, there is no clear information about which attributes from the original dataset are used.

In the ENA dataset, there are two attributes, the gross ENA, and the storable ENA. The gross ENA is the energy generated by the power plant system, which is assumed to operate at 65% of the useful operating level (i.e., the natural water flow into the reservoir), while the storable ENA is equal to the difference between the natural inflow and the flow into the reservoir. The quantity of EAR represents the energy related to the amount of water stored in the reservoir, which can be converted into power generation for the plant itself and for all the plants downstream of the cascade. The maximum ERA represents the storage capacity if the system is full, while the downstream subsystem considers the use of water from the reservoir to generate energy at the downstream power plants, which are in different subsystems.

Theoretically, the representation of the hydropower plant capacity depends on the cascade of hydropower plant turbines in the reservoir. It includes parameters such as the volume of water flowing into the reservoir in the cascade, the type of hydropower plant (run-of-river, reservoir-based, pumped storage), the head height of the plant, the spill, and the volume of water flowing into the reservoir[32]. Typically, only the first hydropower plant has incoming water flow, as downstream plants are constrained by upstream turbine control and spill schedules. However, the hydro station cascades and the mentioned parameter at the resolution of the level of each hydropower plant are not accessible by the authors. Besides, there is no information about the interrelationship between individual hydropower plants with the aggregated level of reservoirs, basins, or REE. Even though ONS publishes the information of the power stations which the basin where they are located, ONS only provides the name of the hydropower plants. We have tried to use string matching to find the exact hydropower station in the previous dataset, but this has not been successful because the naming is not comparable between different datasets. Since the ENA reflects the potential power generation of the hydroelectric power, which is calculated by the volume stored in the reservoirs at their respective operating level, we use this attribute to be the inflow to the hydropower system. To represent the hourly feed-in to the hydropower plants aggregated at the federal state level, we, therefore, use the dataset of ENA spatially resolved by electric region.

Data processing

The ENA data used is daily resolved and is given with a unit MW_month (in PT: MWmês). This is converted to MWh as it is equivalent to the 720 $\mathrm{MWh/month}$[30].

To represent the inflow of the hydropower plants in the federal state, we assume that the inflow of the hydropower plants in each federal state correlates to the installed capacity of the reference year. The installed capacity is the one from the subsection Power plants infrastructure including the hydropower plants of different sizes. Therefore, we get the inflow of the hydropower plants at the aggregated level, in figure 7. Since the installed capacity for a given reference year has two operating states – operating and planning – the inflow dataset provided in this study can be allocated by installed capacity or by the total value of installed and planned capacity.

Variable renewable potentials (wind and solar)

For planning future energy systems, knowledge of the technical generation potential of vRES is essential. In particular, this includes geo-referenced data on the nominal installation capacity that can be installed in a given area, along with an hourly generation time series due to the intermittent generation.

Data collection

The global resource assessment tool, Energy Data Analysis Tool (EnDAT)[33,34], is used for the assessment of renewable energy generation potential of different technologies such as PV, onshore, and offshore wind turbines. Inputs are weather resource maps at an hourly temporal resolution and a spatial resolution of 0.09°×0.09° and static land cover maps, which are provided at a resolution of 0.09°×0.09°. As output, EnDAT provides 1) spatially resolved maximum generation capacity and 2) relative profiles of hourly power feed-in from wind and solar energy. The spatial output resolution of 0.09°×0.09° is aggregated to the level of administrative regions, namely, the federal state level in this paper.

For calculating the installable capacity, one set of maps is used as areas of exclusion (table 15) and another set is serving as suitability criteria to determine the share of the available remaining area (table 16). The spatial land cover maps are based on the Copernicus land cover dataset[35], the global lakes and wetland database[36], IUCN protected area categories[37], and a digital soilmap of the World (for Dunes, Glaciers, Saltpans)[38]. The roughness length is calculated using the land cover maps and a roughness lookup-table[39]. Furthermore, to generate the feed-in time series, we use the spatio-temporal resolved maps from the ERA-5[40] dataset. It contains hourly resolved data for Direct Normal Irradiance (DNI), Global Horizontal Irradiance (GHI), wind speed, and temperature at a 31 km spatial resolution.

Data processing

The maximum installation density is calculated by geometric constraints, i.e. for wind the wake is taken into consideration and for PV the maximal shading during the winter solstice at the assumed module angle is used. It is restricted by the available area, considering information on the land cover of the area, such as bare ground, crops, grasslands, mosses, shrubs, forests, urban area, and roughness, as well as excluded areas such as distance from settlements, elevation, mining sites, protected areas, glaciers, slopes, wetlands, and water depth for offshore winds. By fulfilling any of the exclusion criteria or violating one of the inclusion criteria exclusion masks are created and used to constrain the area that needs to be computed in the potential analysis. The resulting exclusion criteria are provided in table 15.

Next, suitability factors (table 16) are used to obtain the share of area available per land-cover type that can be used to install a particular technology. Therefore, for each power generation technology, a projection of the techno-economical parameters into the year 2050 is performed (table 17). The potential for PV capacities is determined for rooftops, facades, and other surfaces in urban areas and open areas where ground-mounted PV is installed. At the given resolution one pixel can have more than one land cover type and hence the shares of each pixel are considered additive. The resulting installable capacity is an averaged value.

Based on the assessment of maximum generation capacities, the subsequent evaluation of the feed-in time series is performed. Weather data are converted into power generation in each pixel and weighted by the spatial distribution of the installable generation capacities. For PV, feed-in time series are computed based on the module angle, orientation, and the hourly sun position at a temporally resolved GHI, DNI, and temperature profile. The wind feed-in time series takes into account the hourly wind speed (corrected at hub height using the local roughness) and power curves of turbines[33,34]. Finally, generation capacities and time series are spatially aggregated to a defined region – the federal state level in Brazil.

The installable potential map (in $\mathrm{MW/km^2}$) and the annual power production map (in $\mathrm{MWh/km^2}$) illustrate the resource maps obtained from the Brazilian potential analysis. Figure 8 indicates PV generation and figure 9 indicates wind generation, where large geographical features such as bodies of water or rain forests are visible.

Cross-border electricity exchanges

In addition to the national electricity transmission, SIN connects Brazil to Uruguay, Argentina, and Venezuela for importing and exporting electricity to these countries. Annual power imports remain modest, accounting for only 0.04% (0.60 TWh) of total annual energy consumption, with most of the imports happening between May and November.

Data collection

ONS publishes hourly historical cross-border flows with Uruguay and Argentina. We acquire the dataset in July 2021, when the Argentina-Brazil time series is available for the time horizon 1999-2020 and the one for Uruguay-Brazil for 2000-2020.

Data processing

The cross-border power exchange time series from ONS have gaps. In particular, the data for Uruguay-Brazil has one or two missing values for each year except 2018-2020. Most of the data is missing for 2000-2003, and 2.5% of values are missing in 2016 and 0.3% in 2014. The Argentina-Brazil data-set has 1 or 2 missing values in each year except 2019-2020. For 2008, 2009, and 2016 we observe missing shares of missing values of 6.6%, 12.1%, and 2.5% respectively. To be consistent with the time frame of other datasets, only the time series for 2012-2020 are selected to be further processed. Missing values are mainly filled with the value of the same point in time of the previous week. For the rest, the values from the previous hour are used.

The substations of both transmission lines are located in the Rio Grande do Sul (RS) in the Brazilian territory[41]. We manually label the IDs for federal states using two characters, whereas for foreign nodes we use three characters (URU for Uruguay, ARG for Argentina). Accordingly, the transmission liens a labeled as RS-URU and RS-ARG, as illustrated in figure 10.

Data Records

The dataset provided in this paper is available for download from the public repository, `https://doi.org/10.5281/zenodo.6951435`. The download file contains a structure with nine directories, one per subset. Figure 11 illustrates the folder structure. The data files within each directory are in the common format of CSV, except for defined region data. All data are spatially resolved at the ISO 3166-2 level and temporally resolved in hours. The time series files are provided for the reference years from 2012 to 2020.

1 Defined region This folder contains one shapefile which can be opened in geographic information system software. The CRS is EPSG:4087. The description of the records are detailed in table 18. This determines the nodes used for the entire dataset provided in this paper, i.e., the abbreviations of the federal states.

2 Grid network topology The transmission grid data is given in one file, including AC and HVDC transmission lines. Table 19 explains the records. The voltage is not shown here because it is an equivalent network for which the net transfer capacity is calculated.

3 Renewable potentials (wind and solar) The data records are organized in a directory structure containing CSV files. Three generation technologies (wind onshore, wind offshore, PV) are located in three directories: onshore, offshore, and solar_pv.

These directories contain the installable potentials as well as the yearly time series. The installable potentials are named as `EnDAT_<TECH_NAME>_installable_capacity.csv` and the time series as `EnDAT_<TECH_NAME>_per_unit_generation_weather_year_<YYYY>.csv`. The text "YYYY" corresponds to the weather year. The installable capacities contain the region abbreviation and the installable capacity in MW.

The first column of the generation time series data is the hourly timestamp, `YYYY-MM-DD HH:00:00`, the values in the other columns are the unit generation for each federal state and the column names are the abbreviations for each federal state.

4 Installable capacity for biomass thermal plants The subset includes files of the installable capacity records, one per year, each named `biomass_geographic_potential_reference_year_<YYYY>.csv`. The text "YYYY" corresponds to the reference year. Table 20 reports the details of the information provided by each record.

5 Inflow for hydropower plants The inflows to the hydropower plants in each federal state are obtained separately from two allocation parameters related to the operating status of the total installed capacity. Therefore, there are two subdirectories under this folder, namely, `by_hydropower_plants_operation+planning` and `by_hydropower_plants_operation`. Each subdirectory includes nine files for each reference year. The columns of each file specify the federal state, and each row represents the inflow for that federal state throughout the year at `YYYY-MM-DD HH:00:00`, in MWh.

The disaggregated inflows to hydropower plants at each federal state are integrated by two allocation parameters related to the operation status of the installed capacity of total hydropower plants.

6 Power plants infrastructure Table 21 presents the description of the records. The information on the installed capacity of power plants in each federal state is recorded in relation to the reference year so that each file represents a record for each reference year.

7 Electricity load profiles It includes two subdirectories, `by_consumer` and `by_consumption`. This is related to the disaggregation of the original dataset, as presented in the subsection Electricity load profiles. Under each subdirectory, there are sectoral load curves by the hour for each reference year, as detailed in table 22.

8 Scenarios of energy demand We provide energy demand data (XLSX format) aggregated by energy carrier and end-use sector for PNE2050 and COPPE as the attributes of the records are detailed in table 23. Due to legal issues, we can only show the IEA data in figure 6. To speed up the processing of the data, we additionally provide the data in CSV format encoded in 'utf-8'.

9 Cross-border electricity exchanges Under this folder, there is only one file named `international_transmission_RS-URU_RS-ARG_2012-2020_hourly.csv`. It stores cross-border electricity imports and exports between Brazil and its neighbors for the 2012-2020 timeframe. A description of the records on the file is presented in table 24.

Technical Validation

The majority of the original dataset is directly taken from official Brazilian databases and therefore not additionally validated.

One exception is the dataset described in the subsection Variable renewable potentials. The technical validation is approached with two data sources: 1) observations of site-specific power generation from a set of real-world PV plants and wind farms in 2018[42], and 2) country-wide power generation indicators of global databases for 2019: the Global Wind Atlas (GWA)[43] and the Global Solar Atlas (GSA)[44].

Solar feed-in

The spatial distribution of PV plants is shown in figure 4. We gather the installed capacity for each PV park based on[13,45]. Of the 17 PV parks, we use twelve PV parks for further analysis. The Pearson correlation is calculated to indicate whether the temporal profiles generated by simulation and observation are similar. For each PV park shown in table 25, we obtain an average correlation between the simulated data and the reference of approximately 0.8. The deviation can be explained that The orientation and inclination of the reference PV installation cannot be determined from available data sources, such as aerial images. As the effect of the orientation under small module inclinations is minimal, this aspect is not considered in our assessment with EnDAT. By default, EnDAT calculates an ensemble of solar power plants with a southern orientation and facing east and west at 60° away from the south.

The quantity for country-wide validation with the GSA[44] is calculated by converting solar resources into power generated per unit of capacity of pre-defined PV power plants over the long-term, called PVOUT. The solar resources are DNI and GHI obtained by Solargis. EnDAT uses DNI and GHI from ERA5 reanalysis data. Since the GSA provides raster data of 1 km and the spatial resolution of EnDAT is 0.09°, we upscale the GSA data to the spatial resolution of EnDAT using the nearest neighbor method. We use the mean bias error (MBE) to assess the difference in levels of overestimation or underestimation. The comparison of PVOUT derived from EnDAT and the GSA shows an MBE of -7% and -36% PV in urban areas and open areas, respectively. Especially, the deviation increase with the distance to the equator. This indicates that EnDAT overestimates the PVOUT in comparison with GSA. The deviation could be explained by the different solar resource data[46] and is within a reasonable range.

Wind feed-in

The spatial distribution of wind power plants is shown in figure 4 (only onshore wind).

As a validation dataset, we use the observed hourly electricity production in 2018 published by ONS[42] of several wind farms, and gather the installed capacity, hub height, location, and turbine type of each wind farm based on[13,47]. Out of the obtained eleven Brazilian wind farms, we use seven wind farms for further analysis due to data inconsistencies for the rest. As detailed in table 26, the correlation between real-world wind farms and EnDAT-simulated generation time series ranges from 0.23 to 0.58. An explanation for this is the following: The potential analysis approach of EnDAT does not account for local wind effects due to elevation, which may lead to gaps in correlation. However, almost all existing wind farms are located in an area that is highly influenced by local wind phenomena. Several of the investigated wind parks are located on the coastline where hot winds can cause temperature differences between land and sea, superposed with nearby elevation changes inland of the wind parks. Other sites are located on plateaus in hilly terrain. The weather data used by EnDAT – wind speed – originates from the reanalysis data – ERA5, which is of resolution of 31 km at the equator, and the value is represented as the grid average. On tiny geographical and temporal dimensions, however, observations of wind speed can differ due to the local terrain, vegetation, and built environments[40]. The wind speed data from ERA5 rarely accurately describes the wind speed in highlands or valleys.

The data of GWA 3.0[43] is derived from the same reanalysis data as we use, ERA5. However, the GWA only provides average wind speed and average power density at five different heights (10, 50, 100, 150, and 200 m and average capacity factors (CFs) for three turbine classes as defined by the International Electromechanical Commission (IEC). We compare the CFs for IEC class I and III from GWA with simulated results from EnDAT. Since the spatial resolution of GWA is 250 m and the spatial resolution of EnDAT is 0.09° for Brazil, the GWA data is upscaled to the resolution of EnDAT, using the nearest neighbor method. We find in general that our CFs for onshore wind is lower, while they are greater for offshore wind although we adapted our technical specifications to be in line with the assumptions of the GWA (i.e. Vestas V112 for IEC class I and V136 turbines for IEC class III are used for this validation). In particular, the MBE between CFs calculated for onshore wind between EnDAT and the GWA are -17% for IEC Class I and -18% for IEC Class III. The MBE for offshore wind is 17% in IEC Class I and 14% for IEC Class III. Due to barriers in accessing details on assumptions made for the GWA, we are unable to identify all the factors that contribute to the differences between our data processed and the GWA.

Conclusion

To summarize, our simulations correlate better with real-world PV generation than for onshore wind generation at a spatial resolution of 0.09°. Compared to the GSA, EnDAT calculates a higher PV generation than the GSA, however, the onshore wind generation potential is assessed lower compared to the GWA though EnDAT calculates higher offshore wind generation potentials. It is to be emphasized that the data we provide for PV and wind power are aggregated to large geographical areas, i. e. at the federal state level. For this level, appropriate validation data is lacking as both wind and solar are only solely exploited energy resources in the Brazilian power system. Therefore, accessible data for validation is only available for a few locations with partially very individual local peculiarities. Our data shows better agreement with simulated data from other sources – the GSA and GWA, which rely on much higher resolved resources data but only provide CFs instead of time series of power generation. However, downscaling may be necessary when using the regionalized results of EnDAT.

Usage Notes

The dataset provided in this paper consists of several CSV files, so any software that handles CSV files can load the dataset. Most aspects related to the use of this dataset are self-explanatory. The dataset we present here can be used as input to any energy system model, emulating the Brazilian power system in high resolution, hourly, and for the 27 federal states. With this dataset, users are able to perform future scenario analyses. With such a high-resolution harmonized dataset, users can aggregate and represent Brazil in a global energy system model at a sufficient resolution.

However, it is not appropriate to compare historical annual trends in data where the reference year is determined by the installed capacity, for instance, 4 Installable capacity for biomass thermal plant and 5 Inflow for the hydropower plants. This is due to the fact that most of the date information in the original dataset is missing, as described in the subsection Power plants infrastructure.

To achieve the objective of this work, which seeks to provide a reliable and open database to model the power sector in Brazil, we in subsection Scenarios of energy demand make available the evolution of electricity consumption by consumption sector until 2050. For understanding the dynamics of the evolution of electricity consumption, it is necessary to learn the main premises of each scenario, presented in table 13. For example, the dispute between electrification and biofuels (aggregated to total liquids) in the transport sector. Therefore, to better comprehend the role of electrification in each sector and the intersectoral dynamics, the evolution of the consumption of additional energy carriers in each sector until 2050, is also presented in complementary form (cf. figure 6). In addition to this, the PNE2050 data may contain numerical deviations as it is derived from the graph.

Code availability

This paper describes the complete data collection process. The data processing is performed using Python 3.9 and the necessary toolboxes, such as Pandas, Geopandas, etc. We regret that we cannot provide scripts for the entire data processing, especially since the vRES potential data is created by the EnDAT framework, which is in the process of being open-sourced. For those data for which the license is "citation", we have been given permission to redistribute the data after modifying it for the purposes of this paper. We do not, however, have permission to publish their original data. We can therefore make these scripts available on request.

References

1. International Energy Agency. *World Energy Outlook 2021* (International Energy Agency, 2021).
2. Pfenninger, S. *et al.* Opening the black box of energy modelling: Strategies and lessons learned. *Energy Strateg. Rev.* **19**, 63–71, 10.1016/j.esr.2017.12.002 (2018).
3. Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S. & Staffell, I. The importance of open data and software: Is energy research lagging behind? *Energy Policy* **101**, 211–215, 10.1016/j.enpol.2016.11.046 (2017).
4. Jensen, T. V. & Pinson, P. Re-europe, a large-scale dataset for modeling a highly renewable european electricity system. *Sci. Data* **4**, 170175, 10.1038/sdata.2017.175 (2017).
5. Wiese, F. *et al.* Open power system data – frictionless data for electricity system modelling. *Appl. Energy* **236**, 401–409, 10.1016/j.apenergy.2018.11.097 (2019).
6. *Designing a Brazilian energy system model for studying energy planning at high spatial and temporal resolution*, vol. 24 (Energy Proceedings, 2021). https://www.energy-proceedings.org/designing-a-brazilian-energy-system-model-for-studying-energy-planning-at-high-spatial-and-temporal-resolution/.
7. Instituto Brasileiro de Geografia e Estatística. Malha municipal: Br_uf, https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?=&t=downloads (2021).
8. Empresa de Pesquisa Energética. Sistema de informações geográficas do setor energético brasileiro: Linhas de transmissão, https://gisepeprd2.epe.gov.br/WebMapEPE/ (2020).
9. Empresa de Pesquisa Energética. Plano decenal de expansão de energia 2029, https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/Documents/PDE%202029.pdf (2020).
10. Cao, K.-K., Metzdorf, J. & Birbalta, S. Incorporating power transmission bottlenecks into aggregated energy system models. *Sustainability* **10**, 1916, 10.3390/su10061916 (2018).
11. Empresa de Pesquisa Energética. Estudos para a licitação da expansão da transmissão: Análise técnico-econômica e socioambiental de alternativas: Relatório r1, https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-276/topico-525/EPE-DEE-RE-025-2020-rev0+SMA%20-%20Estudo%20para%20Controle%20de%20Tens%C3%A3o%20e%20Suprimento%20ao%20Extremo%20Sul%20da%20Bahia.pdf (2020).
12. Oeding, D. & Oswald, B. R. *Elektrische Kraftwerke und Netze* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

13. Agência Nacional de Energia Elétrica. Sistema de informação de geração da aneel- siga, https://app.powerbi.com/view?r=eyJrIjoiNjc4OGYyYjQtYWM2ZC00YjllLWJlYmEtYzdkNTQ1MTc1NjM2IiwidCI6IjQwZDZmOWI4LWVjYTctNDZhMi05MmQ0LWVhNGU5YzAxNzBlMSIsImMiOjR9 (2021).

14. Operador Nacional do Sistema Elétrico, Empresa de Pesquisa Energética & Câmara de Comercialização de Energia El(étrica. Avaliação e compatibilização das informações de geração, carga e consumo de energia elétrica no sin, https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-251/topico-315/NT_Carga_ONS-EPE-CCEE%20_07-12-2016[1].pdf (2016).

15. Operador Nacional do Sistema Elétrico. Dados da capacidade instalada de geração, https://ons-dl-prod-opendata.s3.amazonaws.com/dataset/capacidade-geracao/DicionarioDados_Capacidade_Instalada_Geracao.pdf (2022).

16. Portugal-Pereira, J., Soria, R., Rathmann, R., Schaeffer, R. & Szklo, A. Agricultural and agro-industrial residues-to-energy: Techno-economic and environmental assessment in brazil. *Biomass Bioenergy* **81**, 521–533, 10.1016/j.biombioe.2015.08.010 (2015).

17. Soria, R. *et al.* Modelling concentrated solar power (csp) in the brazilian energy system: A soft-linked model coupling approach. *Energy* **116**, 265–280, 10.1016/j.energy.2016.09.080 (2016).

18. Operador Nacional do Sistema Elétrico. Histórico da operação da curva de carga horária, http://www.ons.org.br/Paginas/resultados-da-operacao/historico-da-operacao/curva_carga_horaria.aspx (2021).

19. Empresa de Pesquisa Energética. Anuário estatístico de energia elétrica 2021: raw data, https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/anuario-estatistico-de-energia-eletrica (2021).

20. Dranka, G. G. & Ferreira, P. Planning for a renewable future in the brazilian power system. *Energy* **164**, 496–511 (2018).

21. Ministério de Minas e Energia & Empresa de Pesquisa Energética. Plano nacional de energia 2050, https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/Plano-Nacional-de-Energia-2050 (2020).

22. Cao, K.-K., Cebulla, F., Gómez Vilchez, J. J., Mousavi, B. & Prehofer, S. Raising awareness in model-based energy scenario studies—a transparency checklist. *Energy, Sustain. Soc.* **6**, 10.1186/s13705-016-0090-z (2016).

23. Junne, T. *et al.* How to assess the quality and transparency of energy scenarios: Results of a case study. *Energy Strateg. Rev.* **26**, 100380, 10.1016/j.esr.2019.100380 (2019).

24. Baptista, L. B. *et al.* Good practice policies to bridge the emissions gap in key countries. *Glob. Environ. Chang.* **73**, 102472, 10.1016/j.gloenvcha.2022.102472 (2022).

25. van Soest, H. L. *et al.* Global roll-out of comprehensive policy measures may aid in bridging emissions gap. *Nat. communications* **12**, 6419, 10.1038/s41467-021-26595-z (2021).

26. Riahi, K. *et al.* Cost and attainability of meeting stringent climate targets without overshoot. *Nat. Clim. Chang.* **11**, 1063–1069, 10.1038/s41558-021-01215-2 (2021).

27. Killingtveit, Å. Managing global warming: 8 - hydropower. In Trevor M. Letcher (ed.) *Managing Global Warming*, 265–315, 10.1016/B978-0-12-814104-5.00008-9 (Elsevier, 2019).

28. Operador Nacional do Sistema Elétrico. Energia agora: Reservatórios, http://www.ons.org.br/paginas/energia-agora/reservatorios (2022).

29. Operador Nacional do Sistema Elétrico. Energia natural afluente por subsistema, http://www.ons.org.br/Paginas/resultados-da-operacao/historico-da-operacao/energia_afluente_subsistema.aspx (2021).

30. Diuana, F. A., Viviescas, C. & Schaeffer, R. An analysis of the impacts of wind power penetration in the power system of southern brazil. *Energy* **186**, 115869, 10.1016/j.energy.2019.115869 (2019).

31. Fichter, T., Soria, R., Szklo, A., Schaeffer, R. & Lucena, A. F. Assessing the potential role of concentrated solar power (csp) for the northeast power system of brazil using a detailed power system model. *Energy* **121**, 695–715, 10.1016/j.energy.2017.01.012 (2017).

32. Liu, H., Andresen, G. B., Brown, T. & Greiner, M. A high-resolution hydro power time-series model for energy systems analysis: Validated with chinese hydro reservoirs. *MethodsX* **6**, 1370–1378, 10.1016/j.mex.2019.05.024 (2019).

33. Scholz, Y. *Renewable energy based electricity supply at low costs: development of the REMix model and application for Europe*. Ph.D. thesis, Universität Stuttgart, http://dx.doi.org/10.18419/opus-2015 (2012). http://dx.doi.org/10.18419/opus-2015.

34. Daniel Stetter. *Enhancement of the REMix energy system model: Global renewable energy potentials, optimized power plant siting and scenario validation*. Ph.D. thesis, Universität Stuttgart, http://dx.doi.org/10.18419/opus-6855 (2014). http://dx.doi.org/10.18419/opus-6855.

35. Buchhorn, M. *et al.* Copernicus global land service: Land cover 100m: collection 3 : epoch 2015: Globe, https://doi.org/10.5281/zenodo.3939038 (2015).

36. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol.* **296**, 1–22, https://doi.org/10.1016/j.jhydrol.2004.03.028 (2004).

37. Dudley, N. *Guidelines for applying protected area management categories including IUCN WCPA best practice guidance on recognising protected areas and assigning management categories and governance types* (IUCN, 2013).

38. Land and Water Development Division, FAO, Rome. The digital soil map of the world.

39. Silva, J., Ribeiro, C., Guedes, R., Rua, M.-C. & Ulrich, F. Roughness length classification of corine land cover classes. *Proc. EWEC 2007* (2007).

40. Hersbach, H. *et al.* The era5 global reanalysis. *Q. J. Royal Meteorol. Soc.* **146**, 1999–2049, https://doi.org/10.1002/qj.3803 (2020). https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

41. Operador Nacional do Sistema Elétrico. Mapa do sistema de transmissao: Horizonte 2024, http://www.ons.org.br/_layouts/download.aspx?SourceUrl=http://www.ons.org.br/Mapas/Mapa%20Sistema%20de%20Transmissao%20-%20Horizonte%202024.pdf (2021).

42. Operador Nacional do Sistema Elétrico. Histórico da operação da geração de energia, http://www.ons.org.br/Paginas/resultados-da-operacao/historico-da-operacao/geracao_energia.aspx (2020).

43. Badger, J. *et al.* Global wind atlas 3.0. *Last accessed* (2019).

44. Solargis. Global solar atlas 2.0: A free web-based application developed and operated by the company solargis s.r.o. on behalf of the world bank group, utilizing solargis data, with funding provided by the energy sector management assistance program (esmap). *Last accessed* (2019).

45. Operador Nacional do Sistema Elétrico. Boletim mensal de geração solar, http://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/Boletim%20Mensal%20de%20Gera%C3%A7%C3%A3o%20Solar%202021-04.pdf (2021).

46. Urraca, R. *et al.* Evaluation of global horizontal irradiance estimates from era5 and cosmo-rea6 reanalyses using ground and satellite-based data. *Sol. Energy* **164**, 339–354, https://doi.org/10.1016/j.solener.2018.02.059 (2018).

47. Operador Nacional do Sistema Elétrico. Boletim mensal de geração eólica, http://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/Boletim%20Mensal%20de%20Gera%C3%A7%C3%A3o%20E%C3%B3lica%202021-02.pdf (2021).

48. Operador Nacional do Sistema Elétrico. Histórico da operação: Intercâmbios de energia, http://www.ons.org.br/Paginas/resultados-da-operacao/historico-da-operacao/intercambios_energia.aspx (2021).

49. Rodrigues, E., Pontes, R., Bandeira, J. & Aguiar, V. Analysis of the incidence of direct lightning over a hvdc transmission line through efd model. *Energies* **12**, 555, 10.3390/en12030555 (2019).

50. Graham, J., Holmgren, T., Fischer, P. & Shore, N. L. (eds.). *The Rio Madeira HVDC System – Design aspects of Bipole 1 and the connector to Acre-Rondônia* (2012).

51. Esmeraldo, P. C. V. Technical benefits of hvdc lines and experience of technical benefits of hvdc lines and experience of hvdc projects: Brazil, china, and perspectives in latin america, https://energia.gob.cl/sites/default/files/mini-sitio/07_stategrid_paulo_esmeraldo.pdf (2020).

52. *The Garabi 2000 MW Interconnection Back-to-Back HVDC to connect weak ac systems* (Citeseer, 2012). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.635&rep=rep1&type=pdf.

53. International Energy Agency. World energy model documentation, https://iea.blob.core.windows.net/assets/932ea201-0972-4231-8d81-356300e9fc43/WEM_Documentation_WEO2021.pdf (2021).

54. Masson-Delmotte, V., P. Zhai, A. Pirani, S.L.Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K.Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Summary for policymakers: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. Cambridge University Press, https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM.pdf (2021).

Acknowledgements

The authors thank Josua Braun for his assistance in compiling the GWA and GSA data for the chapter on technical validation of the vRES potential. The research underlying this work is supported by Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) within a framework founded by the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) through the Klimaneutrale Alternative Kraftstoffe (ProQR) project.

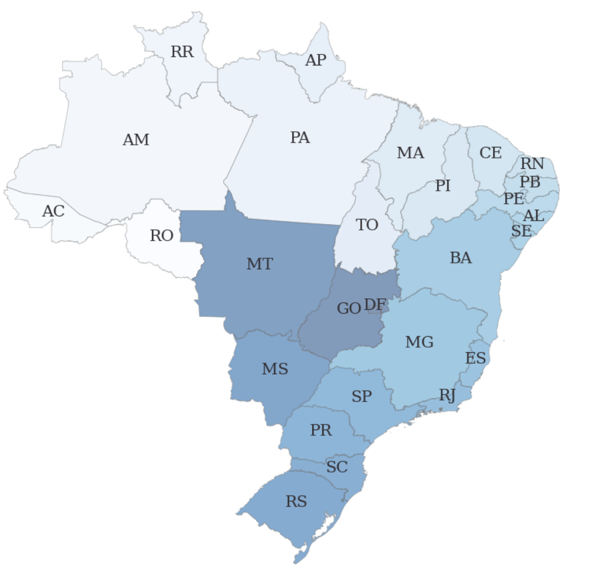
Author contributions statement

Y.D. collected and analyzed data, coded and documented datasets, and wrote the manuscript. R.S., H. W.X, and K.v.K. prepared the dataset and the technical validation of the vRES potential and participated in writing this section of the manuscript. R.S. was involved in the energy demand scenario section of the manuscript. P.R.R.R. provided COPPE scenario data. K.K.C. and P.J. provided scientific guidance and supervision throughout the writing process. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables



Code	Name	Code	Name
AC	Acre	PB	Paraíba
AL	Alagoas	PR	Paraná
AP	Amapá	PE	Pernambuco
AM	Amazonas	PI	Piauí
BA	Bahia	RJ	Rio de Janeiro
CE	Ceará	RN	Rio Grande do Norte
DF	Distrito Federal	RS	Rio Grande do Sul
ES	Espírito Santo	RO	Rondônia
GO	Goiás	RR	Roraima
MA	Maranhão	SC	Santa Catarina
MT	Mato Grosso	SP	São Paulo
MS	Mato Grosso do Sul	SE	Sergipe
MG	Minas Gerais	TO	Tocantins
PA	Pará		

Figure 1. 27 regions defined according to ISO 3166-2 – Brazilian federal states – used in this study

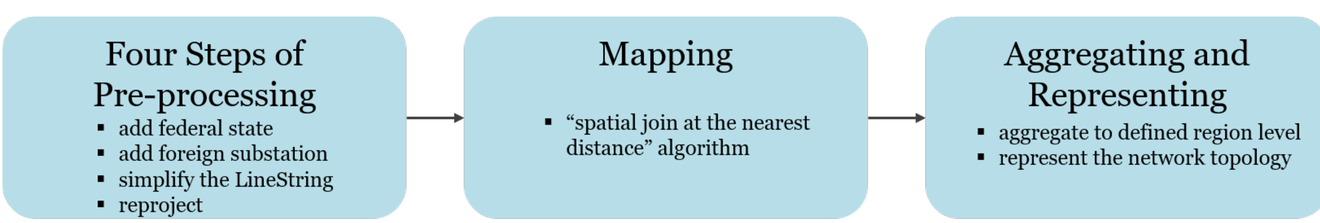


Figure 2. Overview of processing grid network data

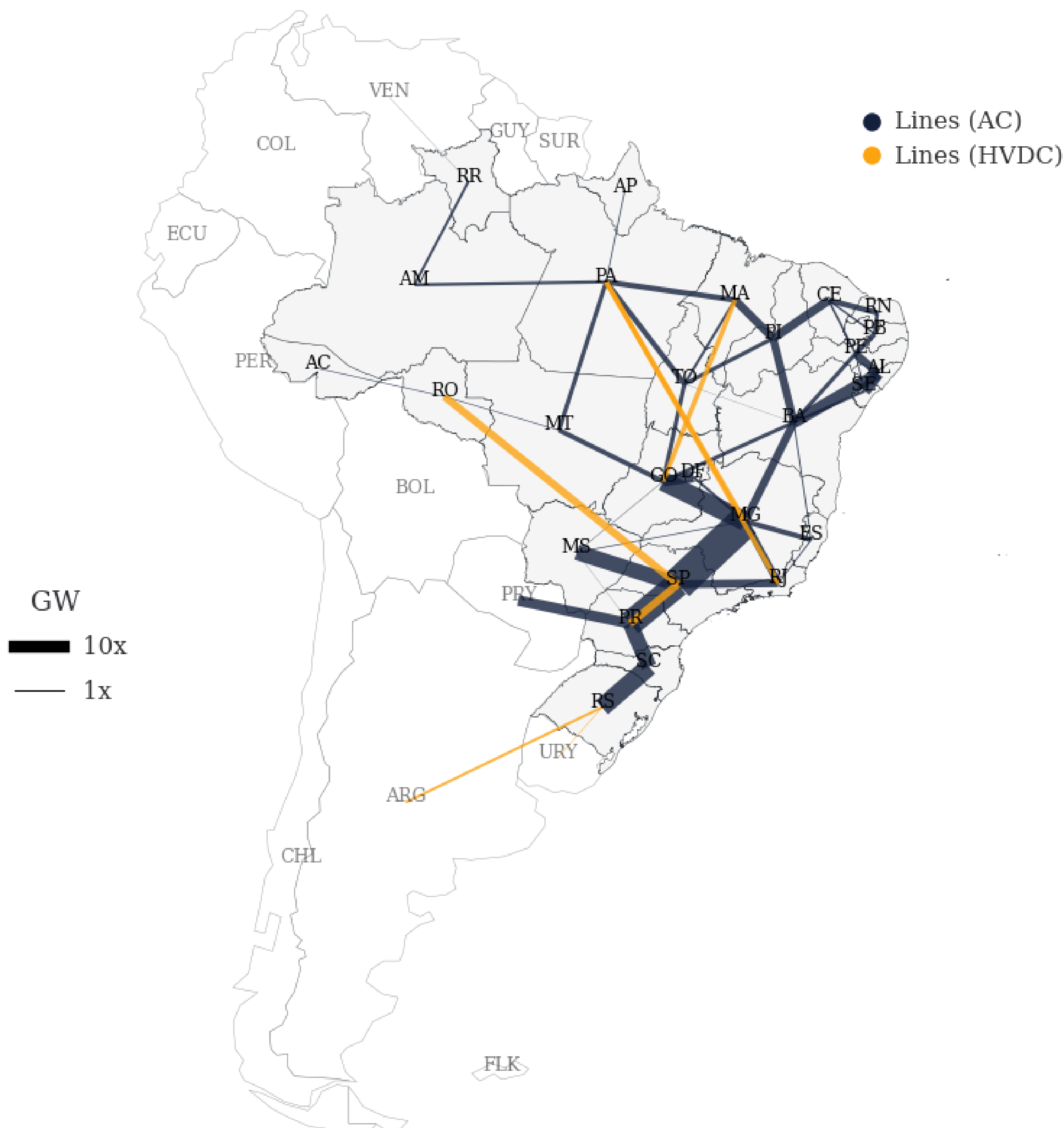


Figure 3. Transfer capacity between the defined regions, GW

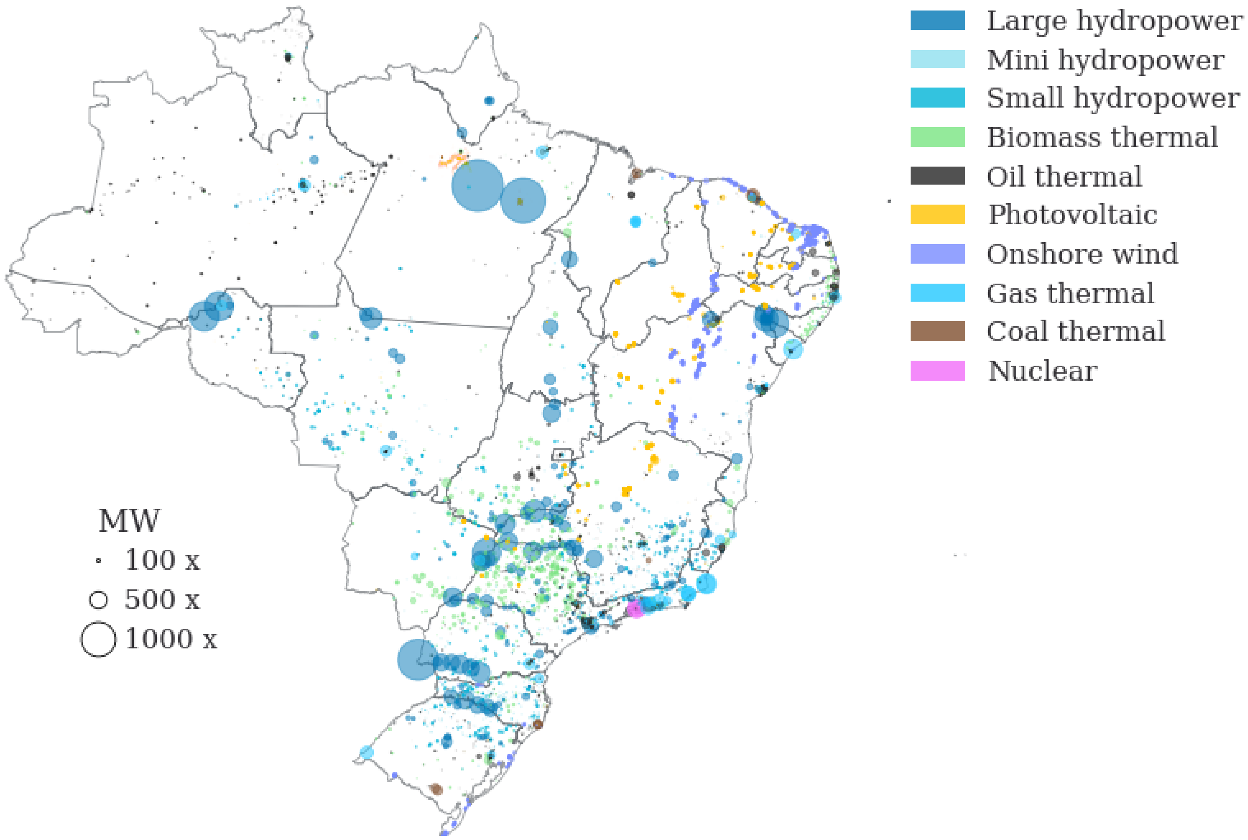


Figure 4. Power plants distribution

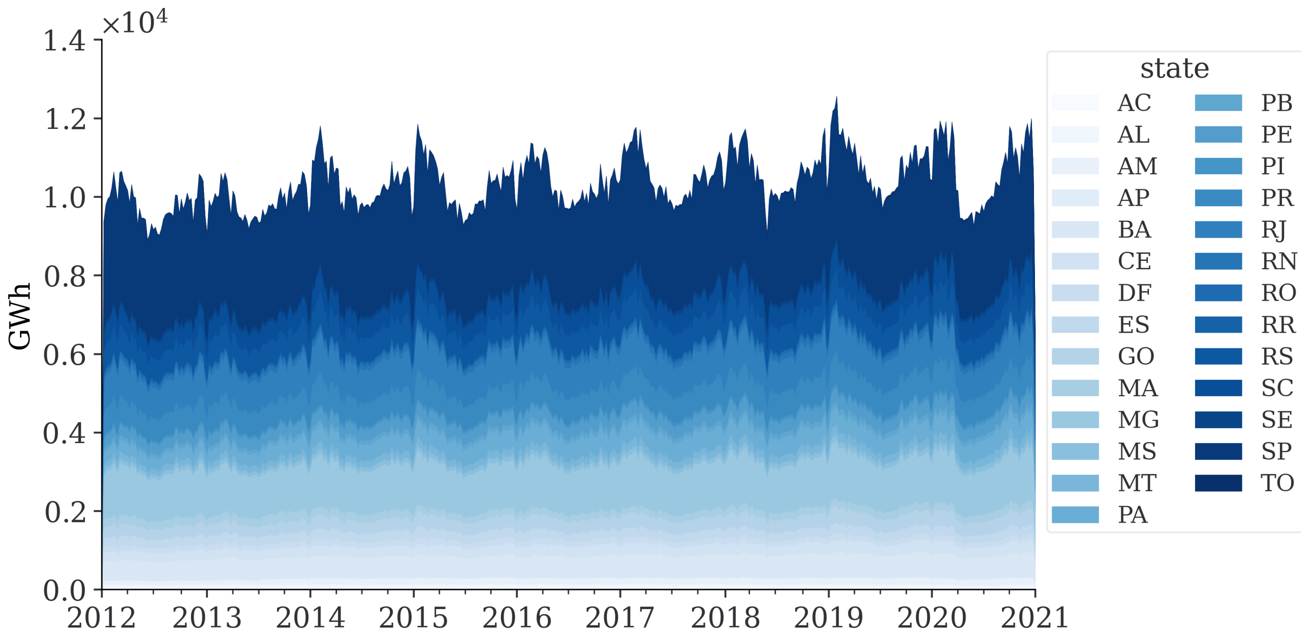
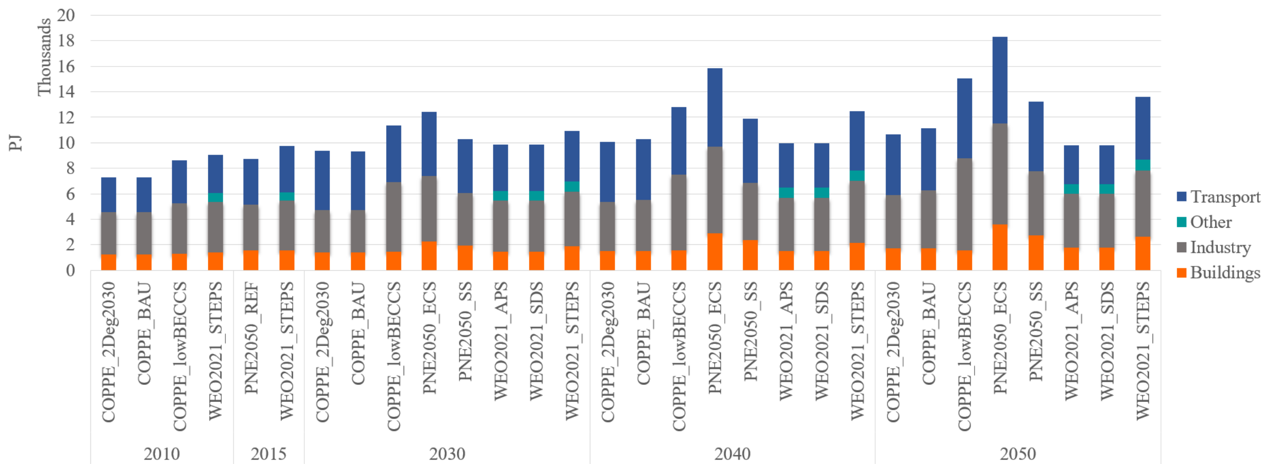
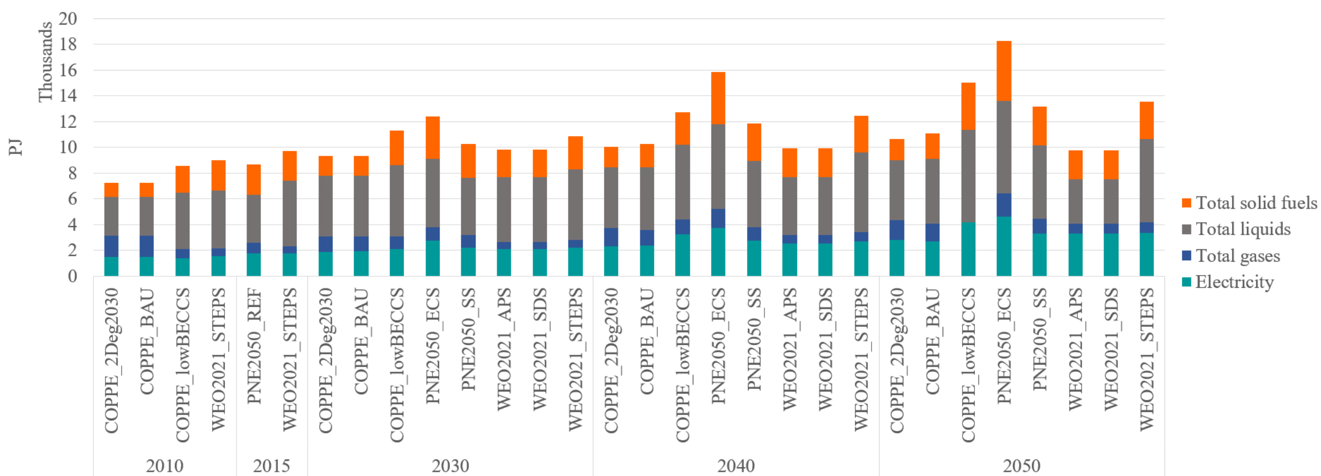


Figure 5. Electricity load at federal states (weekly) distributed by annual consumption. The dataset to be published is hourly resolved



(a) by sector



(b) by energy carrier

Figure 6. Comparison between demand scenarios by different studies

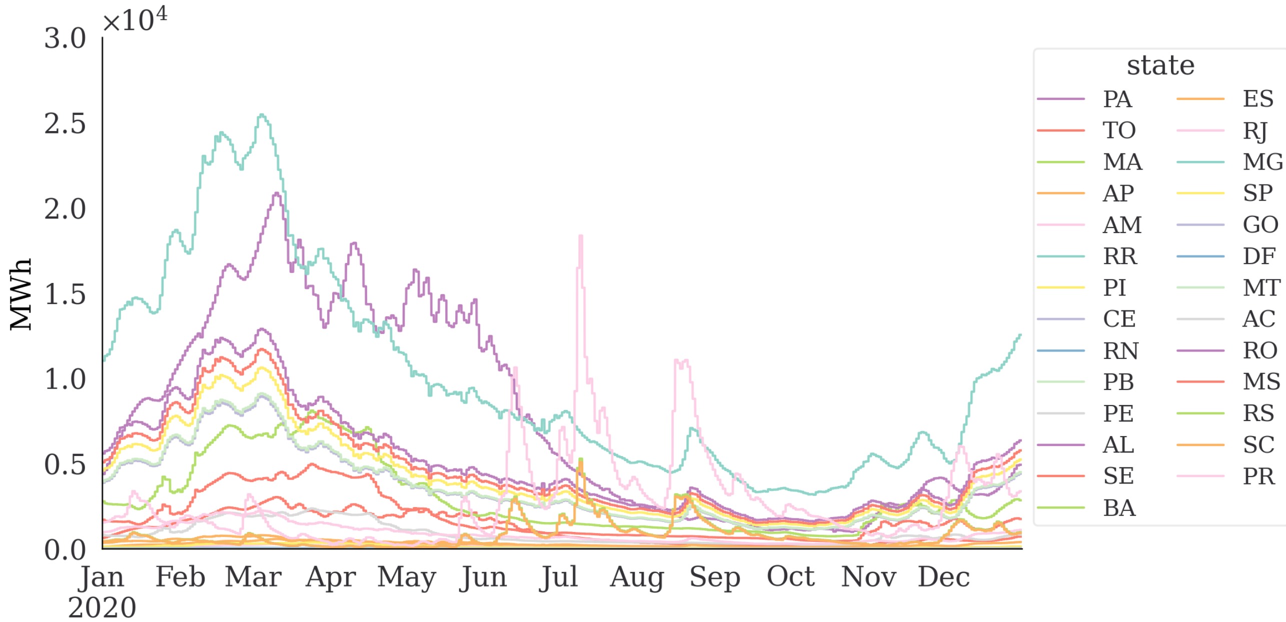
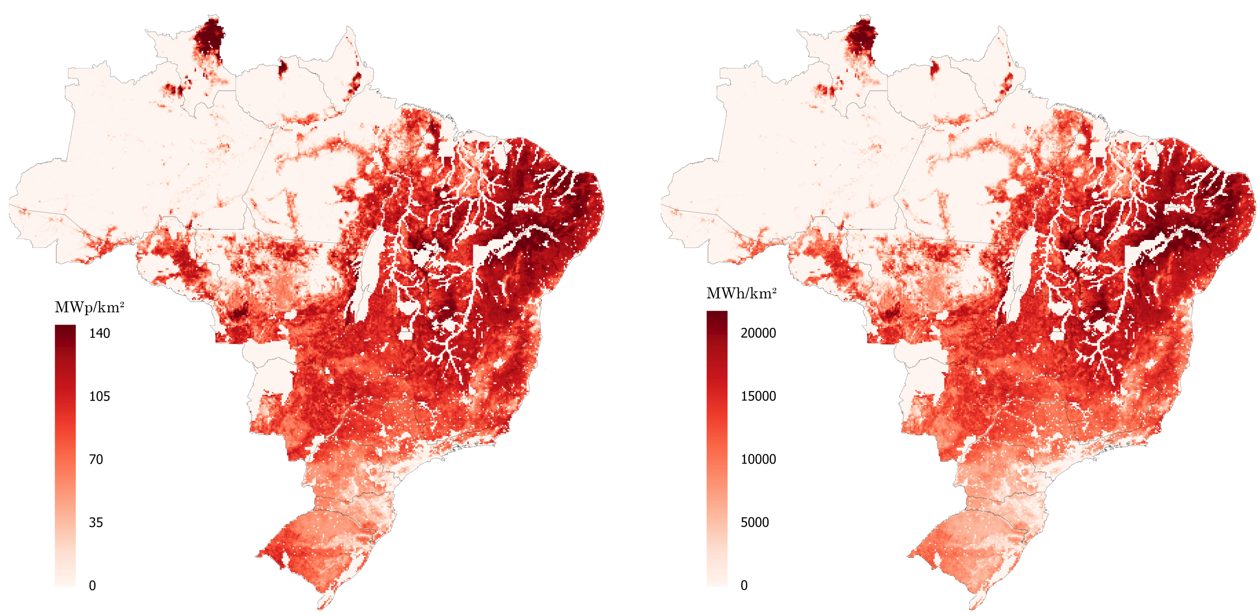


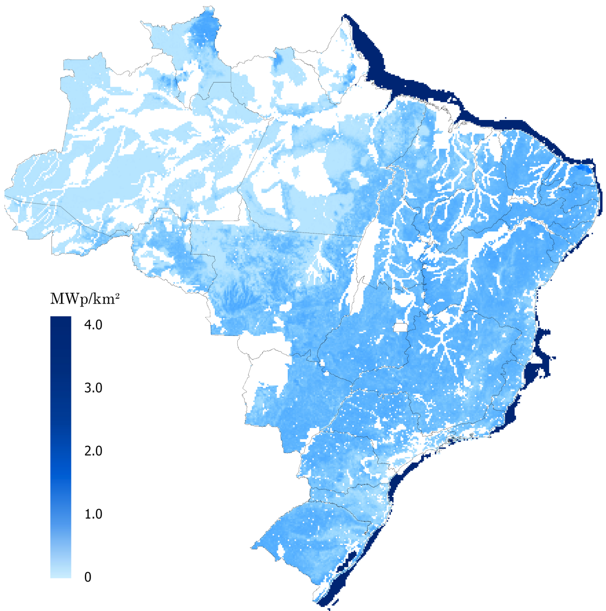
Figure 7. Per federal state inflow for the reference year 2020. Note: allocation is based on the installed capacity (phase is operation) of hydropower plants in the reference year



(a) Installable capacities

(b) Annual generation

Figure 8. Maps of PV power generation potential for the reference year 2019. Each map combines the potential for urban and open field installation and generation.



(a) Installable capacities **(b)** Annual generation

Figure 9. Maps of wind generation potential for the reference year 2019. Onshore wind and offshore wind are combined in each map.

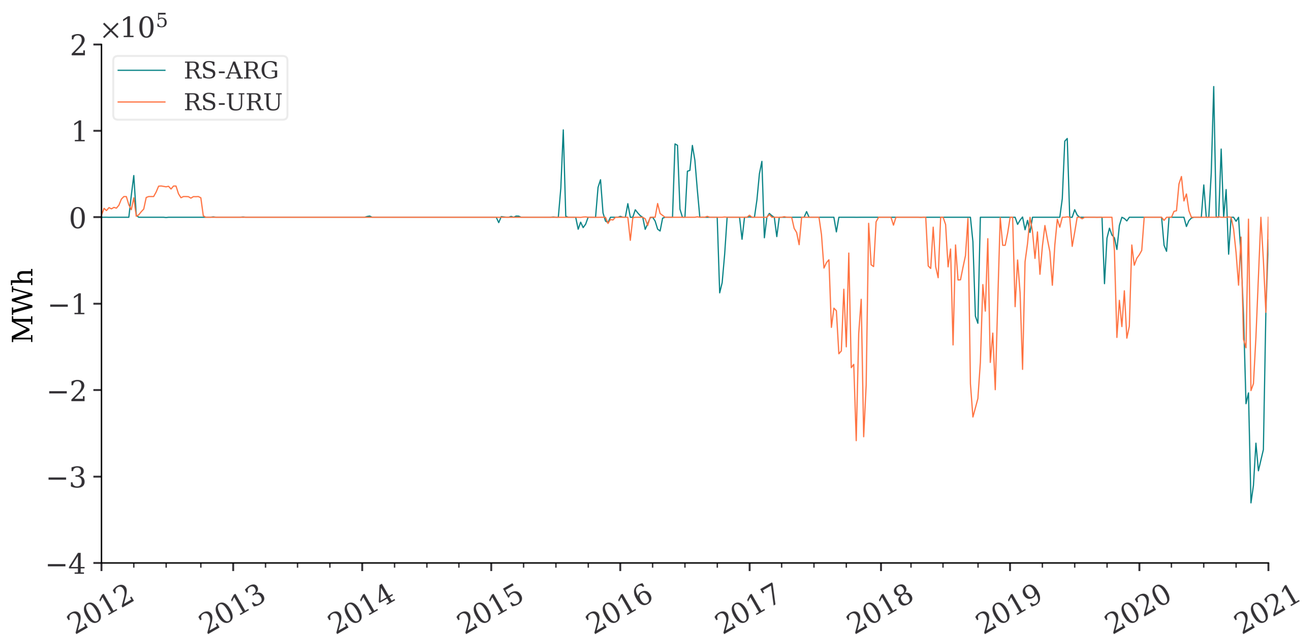


Figure 10. Weekly cross-border electricity transmission from Brazil-Uruguay (RS-URU) and Brazil-Argentina (RS-ARG). Note: The dataset provided is hourly.

Subsection	Raw Dataset	Data License
Defined region	IBGE - Municipal Mesh Data[7]	ODbL [i]
Grid network topology	EPE Webmap[8]	ODbL
Renewable potentials (wind and solar)	[35–40] [ii]	citations [iii]
Installable capacity for biomass thermal plant	[16]	citations
Inflow for the hydropower plants	ONS – Historical Natural Energy Inflow for Each Region[29]	ODbL
Power plants infrastructure	ANEEL-SIGA[13]	ODbL
Electricity load profiles	ONS – Historical Regional Load Curve[18], EPE – Statistical Yearbook of Electricity[19]	ODbL
Scenarios of electricity demand	[1,21,24–26]	citations
Cross-border electricity exchanges	ONS – Historical Energy Exchanges[48]	ODbL

Table 1. Summary of original open accessible datasets used in this paper and their license

[i] Open Data Commons Open Database License (ODbL): https://opendatacommons.org/licenses/odbl/1-0/. Law NO. 8777, of May 11th, 2016 establishes the Open Data Policy to guide the federal executive branch in releasing open government data. The main parts of the policy are the Open Data Plan (PDA, in Portuguese, Plano de Dados Abertos) by each federal agency and a Brazilian Open Data Portal. These datasets have an Open Data Commons Open Database License (ODbL), for example, "Geometric mesh of Brazilian municipalities".

- See: "CAPÍTULO I, Art. 3º -IV permissão irrestrita de reuso das bases de dados publicadas em formato aberto" is "unrestricted permission to reuse databases published in open format".
- See: "CAPÍTULO II, Art. 4º Os dados disponibilizados pelo Poder Executivo federal e as informações de transparência ativa são de livre utilização pelos Poderes Públicos e pela sociedade. (Redação dada pelo Decreto nº 9.903, de 2019)" is "the data provided by the Federal Executive Branch and the information of active transparency are freely used by the Public Authorities and society. (Writing by Decree No. 9,903, 2019)."

[ii] The data is generated by the EnDAT[33,34]. EnDAT is under the procedure of being open source. The dataset published in this paper is the first Brazilian dataset produced by EnDAT.

[iii] We have been given permission to republish the data after modifications made for the purposes of this paper, but this does not mean that we can make their original data publicly available.

Electric regions in SIN	Federal states
North (N)	Pará, Tocantins, Maranhão, Amapá, Amazonas, Roraima
Northeast (NE)	Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia
Southeast/Midwest (SE)	Espírito Santo, Rio de Janeiro, Minas Gerais, São Paulo, Goiás, Distrito Federal, Mato Grosso, Mato Grosso do Sul, Acre, Rondônia
South (S)	Rio Grande do Sul, Santa Catarina, Paraná

Table 2. Electrical regions defined in the SIN and the federal states covered

```
—1 Defined region
|– 2 Grid network topology
|– 3 Renewable potentials (wind and solar)
   |– onshore
   |– offshore
   |– PV
|– 4 Installable capacity for biomass thermal plant
|– 5 Inflow for the hydropower plants
   |– by_hydropower_plants_operation
   |– by_hydropower_plants_operation+planning
|– 6 Power plants infrastructure
|– 7 Electricity load profiles
   |– by_consumption
   |– by_consumer
|– 8 Scenarios of energy demand
└– 9 Cross-border electricity exchanges
```

Figure 11. Folder structure of data records

	Existing	**Planing**	**Total**
Number of lines	1589 [i]	453	2042
Number of substations	735	165	900
Number of power plants	2904 [ii]	274	3178

Table 3. Number of the lines, substations, and power plant units in the original dataset [iii]

[i]The majority of them are at 230 kV and 500 kV, whereas the lowest voltage level is 69 kV. There are three 800 kV lines. 115 lines have a length of less than 1 km and seven lines have a length of more than 2000 km.

[ii]It includes hydropower plants, wind onshore, PV, fossil-thermal, biomass-thermal, and nuclear power plants.

[iii]These are the statistics downloaded by the author during April 2021. The statistics may differ slightly from the version as the data is updated from time to time.

	to sub_0	**to sub_1**
count	2402	2402
mean	1.3	1.4
std	2	3
min	1	1
50%	1	1
90%	1	1
95%	1	2
max	4.8	8.2

Table 4. Statistical summary of the mapping – distance to sub_0 and to sub_1, km

Line name [i]	Transfer capacity	Source
LT 600 kV Foz do Iguaçu – Ibiúna C1	3150	[49]
LT 600 kV Foz do Iguaçu – Ibiúna C2	3150	[49]
LT 600 kV Coletora Porto Velho – Araraquara, C1/C2	3150	[50]
LT 600 kV Coletora Porto Velho – Araraquara, C3/C4	3150	[50]
LT 230 kV Coletora Porto Velho – Porto Velho, C1	400	[50]
LT 230 kV Coletora Porto Velho – Porto Velho, C2	400	[50]
LT 800 kV CC Xingu - Estreito	4000	[51]
LT 800 kV CC Xingu – Terminal Rio	4000	[51]
LT 800 kV CC Graça Aranha – Silvânia	4000	[51]
LT 500 kV Rincón de Santa Maria – Garabi I C1	1100	[52]
LT 500 kV Rincón de Santa Maria – Garabi II C1	1100	[52]
LT 230 kV Livramento 2 – Rivera C1	70	[9]
LT 500 kV Candiota – Melo C1	500	[9]

Table 5. Transfer capacity of HVDC lines added manually, MW

[i] the same name in the original dataset

<table>
<tr><th>ONS Historical Database</th><th>EPE Webmap</th><th>ANEEL-SIGA</th></tr>
<tr><td>solar_pv</td><td>solar_pv</td><td>solar_pv</td></tr>
<tr><td>on_wind</td><td>on_wind</td><td>on_wind</td></tr>
<tr><td>nuclear</td><td>nuclear</td><td>nuclear</td></tr>
<tr><td rowspan="2">thermal</td><td>biomass_thermal</td><td rowspan="2">thermal</td></tr>
<tr><td>fossil_thermal</td></tr>
<tr><td rowspan="4">hydro</td><td>small_hydro</td><td>small_hydro</td></tr>
<tr><td>mini_hydro</td><td>mini_hydro</td></tr>
<tr><td>hydro</td><td>hydro</td></tr>
<tr><td></td><td>wave</td></tr>
</table>

Table 6. The types of power plants used in the three datasets

Type	EPE Webmap	ANEEL-SIGA	ONS Historical Database
hydro	1351	1275	153
nuclear	2	2	2
on_wind	600	506	4
solar_pv	63	924	3
thermal	666	2964	103
total	2682	5671	265

(a) Number of plant units

Type	EPE Webmap	ANEEL-SIGA	ONS Historical Database
hydro	108.08	107.57	109.21
nuclear	1.99	1.99	1.99
on_wind	15.07	12.50	14.27
solar_pv	1.82	1.01	1.78
thermal	36.76	41.32	35.78

(b) Installed capacity per plant type defined by ONS, GW

Table 7. Comparison plant units in operation for the reference year 2018 included between EPE Webmap, ANEEL-SIGA, and ONS Historical Database

Plant type	SIN region	ONS Historical Database	ANEEL-SIGA	EPE Webmap	$\Delta_{\text{ANEEL-SIGA}}$ [i]	$\Delta_{\text{EPE Webmap}}$ [ii]
hydro	N	18.58	24.03	26.54	29.30	42.79
	NE	11.02	11.10	11.03	0.72	0.08
	S	16.60	24.63	23.82	48.40	43.52
	SE	63.00	47.80	46.68	-24.12	-25.91
nuclear	SE	1.99	1.99	1.99	0.00	0.00
on_wind	N	0.22	0.22	0.43	0.01	92.94
	NE	12.00	10.17	12.54	-15.28	4.46
	S	2.02	2.08	2.08	3.25	3.05
	SE	0.03	0.03	0.03	0.56	0.56
solar_pv	N	0.00	0.01	0.00	-	-
	NE	1.24	0.72	1.28	-41.55	3.14
	S	0.00 [iii]	0.01	0.00	36.29	-100.00
	SE	0.54	0.27	0.54	-49.12	0.86
thermal	N	3.93	4.58	5.41	16.42	37.45
	NE	7.35	7.53	5.98	2.45	-18.60
	S	3.87	5.15	4.17	33.15	7.87
	SE	20.63	24.06	21.20	16.64	2.76

Table 8. Comparison at electric region level and harmonized plant type for reference year 2018. The first three columns are in GW, while the rest are in %. Deviations higher than 10% are marked in orange.

[i] $\Delta_{\text{ANEEL-SIGA}} = (\text{ANEEL-SIGA} - \text{ONS})/\text{ONS} * 100$, where the ONS refers to the data from ONS Historical Database

[ii] $\Delta_{\text{EPE Webmap}} = (\text{EPE Webmap} - \text{ONS})/\text{ONS} * 100$

[iii] The value is 3.998 MW. which is rounded to zero when converting the unit to GW.

GGG.FF.UF.999999-D	
Part	**Explanation**
GGG	generation Type
FF	the fuel type abbreviation
UF	federal state abbreviation
999999-D	unique number with identification digit

Table 9. CEG definition

Name	Abbreviation [i]	Explanation
Large hydropower plant	UHE	The hydropower plant with a capacity greater than 5MW and less than 50MW without those identified as small hydro
Small hydropower plant	PCH	The hydropower plant with a capacity greater than 5MW and less than or equal to 30MW with a reservoir area of up to 13 km^2
Mini hydropower plant	CGH	The hydropower plant with a capacity of 5MW or less
Wave power plant	CGU	The energy comes from the water dynamics obtained from the sea waves. The energy comes from the kinetic energy of water from ocean waves. There is only one and the first wave power plant in Latin America, Porto do Pecém with 0.05MW in Ceará.
Thermoelectric plant	UTE	They generate energy with electricity released from any product that generates heat, such as bagasse from various plants, wood chips, fuel oil, diesel, natural gas, enriched uranium, and natural coal.
Thermonuclear plants	UTN	Thermoelectric power plants, using the energy released by nuclear fission of uranium as a source
Wind power plant	EOL	Converting the kinetic energy of the wind into electrical energy. So far, EOL refers to onshore wind power plants
PV power plant	UFV	These are power plants that convert the sun's energy into electricity through the photovoltaic effect, which is a voltage or corresponding current produced by a material when it is exposed to light.

Table 10. Power plants description and their abbreviations

[i] in Portuguese

Year	ONS	EPE	Δ
2012	511.7	448.1	63.6
2013	514.7	463.1	51.5
2014	539.5	474.8	64.7
2015	537.6	465.7	71.9
2016	541.5	461.8	79.7
2017	549.1	467.2	82.0
2018	554.3	474.8	79.5
2019	565.7	482.2	83.5
2020	557.2	475.6	81,5

(a) TWh. $\Delta = \text{ONS} - \text{EPE}$

Year	N	NE	S	SE
2012	-4%	17%	11%	14%
2013	7%	16%	10%	9%
2014	17%	14%	9%	12%
2015	20%	16%	10%	13%
2016	21%	17%	12%	14%
2017	22%	18%	12%	14%
2018	24%	19%	11%	13%
2019	25%	18%	11%	14%
2020	22%	19%	11%	13%

(b) Difference at electric region level

Table 11. Annual electricity consumption between ONS and EPE datasets

Aggregated	**Original in EPE**
Buildings	residential, commercial. public sector, public lighting, public service, own-use
Transport	—
Industry	industrial
Others	rural

Table 12. End-use sector for aggregation and original EPE dataset

Study	IEA WEO2021			
Scope	i) global and regional outlooks, ii) environmental impacts of energy use on emissions and pollutants, iii) impact of policy actions and technological change, iv) fuel supply chain investment requirements to meet projected energy demand, v) prospects of modern energy access			
Scenario	Net Zero Emissions by 2050 Scenario (NZE) [i]	Announced Pledges Scenario (APS) [ii]	Stated Policies Scenario (STEPS) [ii]	Sustainable Development Scenario (SDS)
	“well below 1.5 °C” [iii], an achievable pathway for the global energy sector to achieve net CO_2 emissions by 2050, meeting key energy-related UN Sustainable Development Goals (SDGs). Fully dependent on the emission reductions of the energy sector.	all climate commitments made by governments prior to XXX will be met in full and on time, such as nationally determined contributions (NDCs) and longer-term net-zero goals, objectives, intentions	sector-by-sector assessment to extrapolate current or developing policies and measures of governments around the world	"well Below 2°C" [iii] integrated scenario accomplishes the main UN Sustainable Development Goals related to energy, namely: i) universal access to affordable, reliable, sustainable, and modern energy services by 2030 (SDG 7), ii) significant reduction in air pollution (SDG 3.9), and ii) effective action to mitigate climate change (SDG 13) in order to reach net zero global emissions by 2070.
Narrative	For Brazil, the macroeconomic and demographic assumptions used in all scenarios are different. Energy demand forecasts are based on the average retail price of each fuel used in the end-use, generation, and other conversion sectors. End-use prices are derived from the projected international prices of fossil fuels and subsidy/tax levels. The price of CO_2 is different under different scenarios. However, the narrative for each end-use sector in Brazil cannot be described intuitively.			
Model	A hybrid model – the IEA’s World Energy Model (WEM) of the data-intensive global energy system and the Energy Technology Perspectives (ETP) model of technical and economic parameters of energy technologies. Total final energy demand is the sum of energy consumption in each end-use sector. Energy consumption is assessed with considerable sectoral and end-use detail derived from historical data on the inventories of existing energy infrastructure and socioeconomic variables. For example, the number of vehicles in the transportation sector, the production capacity in the industrial sector, and the floor space in the building sector. The demand for energy services is modeled specifically for each sector. For example, in the residential sector, the demand is divided into space heating, water heating, cooking, lighting, appliances, and space cooling.			
Key drivers	Population: • Annual growth rate (2010-50): 0.3% • Population in 2050: 229 million • Urbanization (% of population) in 2050: 92% GDP: • Average growth rate (2020-50): 2.6% Remaining fossil fuel sources • Oil • Nature gas • Coal Fossil fuel prices: • Nature gas • Steam coal Carbon price End-use prices • Fuel end-use prices • Electricity end-use prices • Wholesale electricity price Derivation of a simplified merit order for thermal power plants Calculation of average marginal cost in each merit order segment Calculation of wholesale price based on average marginal cost Subsidies to fossil fuels			
Source	[1,53]			

(a) WEO21

[i] Global analysis only, not regional analysis
[ii] No specific result designed to be achieved
[iii] “Well below 1.5 °C” and “well below 2 °C”: goals announced in Paris Agreement, limiting the average global temperature increase by 2100 to 2 °C above pre-industrial levels without temperature overshoot (50% probability)

Study	**EPE PNE2050**	
Scope	Brazil's national long-term strategy - National Energy Plan (PNE) 2050 - published by Energy Research Office (EPE), outlines the government's strategic long-term vision in an integrated way.	
Scenario	Expansion Challenge scenarios (ECS)	Stagnation scenarios (SS)
	It reflects the strong growth in total energy demand - a national average annual GDP growth rate of 3.0%. On average, the total final energy consumption increases by 2.2% per year from 2015 to 2050, with faster growth (2.5%) in the first 15 years until 2050, more than twice the 2015 consumption.	It reflects a trajectory that keeps per capita energy consumption around 2015 levels - with a national average annual GDP growth rate of 1.6%. From 2015 to 2050, total energy consumption grows by an average of 1.4% per year (over 10% in total).
Narrative	Consider a more stable economic, political, institutional, and social environment that allows for the completion of important structural reforms that will have a significant impact on the business environment, investment, and productivity, thereby contributing to GDP growth. Overall, total final energy demand is rising, with a decrease in the share of petroleum products and an increase in the share of electricity being driven by all sectors, but mainly by the residential sector.	
Model	The energy demand forecasting methodology consists of three modules - economics study, demand projection, and demand integration. The assumptions used in the specific models for each sector are derived from a discussion of the main moderators and key uncertainties in industry, agriculture and livestock, buildings, services, transport, and meeting electricity and fuel consumption. Long-term economic scenarios are used as one of its main informational inputs. Through the elaboration of sectoral scenarios for agriculture and livestock, services and industrial activities, as well as for infrastructure and mobility needs, the sector-specific model allows for estimating changes in demand by source and by sector at the national level.	
Key drivers	Population: • Annual growth rate: 0.3% • Population in 2050: 226 million • Urbanization in 2015: 86% • Urbanization in 2050: 89% GDP: • Annual GDP growth rate (2016-50): 3.0% • GDP per capita: 2.8% Number of households: • inhabitants/household (2015): 3.2 • inhabitants/household (2050): 2.3 • households (2015): 33 million • households (2050): 98 million	
Source	[21]	

(b) PNE2050

Study	COPPE		
Scope	Integrated long-term scenarios for Brazil from 2010 to 2050 or 2100, in 5-year time steps, using representative time slices. Analyze the competition between technologies and energy sources to meet in a cost-efficient way the demand for energy services under policy and emission targets (modeling explicitly includes the industrial, energy, transportation, residential and commercial, and agricultural sectors)		
Scenario	Business as usual (BAU)	2Deg2030	lowBECCS
	It is based on most likely socioeconomic assumptions of the second marker baseline scenario from the Shared Socioeconomic Pathways (SSP2) throughout the century, with no additional climate policies after 2010.	It is a mitigation scenario consistent with an increase of global warming up to 2 °C without overshooting above pre-industrial levels to 2100. It builds on submitted NDCs actions (unconditional and conditional) up to 2030, and then it transitions (constrained by a national carbon budget) in a cost-effective way toward a 2 °C pathway. To model the scenario, a national (2011 – 2050 accumulated) budget of 22 Gt of CO_2 [i], derived from a global budget of 1000 Gt of CO_2 in the period from 2011 to 2100 is used [ii]. This scenario uses the same socioeconomic assumptions as the BAU scenario.	It is a mitigation scenario consistent with an increase in global warming up to 1.5 °C above pre-industrial levels in 2100. The "lowBECCSS" is an "End-of-century budget" scenario. In the near-term (2020-2030) it builds on immediate action following implemented national policies (NPi) as of 2020. In the long-term, the CO_2 pathway is constrained by cumulative CO_2 emissions over the entire century, allowing high-temperature overshoot and global net-negative CO_2 emissions (NNCE) in the second half of the century. A global budget of 400 Gt of CO_2 between 2018 and 2100 is used. This scenario is conservative regarding the role of bioenergy with carbon capture and storage (BECCS) on a global scale. To only consider the sustainable global BECCS potential, this technology is capped to around 8 $GtCO_2$/yr in 2100. This scenario incorporates the middle-of-the-road socioeconomic conditions throughout the century, based on the SSP2.
Narrative	These are scenarios used for scientific purposes, which can inform decision-making in the Brazilian Government and globally. The useful energy demands for each consumption sector, the food demands, and the reforestation and deforestation scenarios used are calculated exogenously. For the short-term modeling, official public data is considered, while for the long term, international scenarios are used as a reference. The three scenarios shown contain the same population growth and socioeconomic development trajectory (SSP2), but are subject to different levels of ambition to achieve global climate goals.		
Model	It is calculated during the COMMIT project with the BLUES model. [iii]		This scenario is calculated with the COFFEE v1.1 global model during ENGAGE project. [iv]
Key drivers	Population: • Annual growth rate(2015-50): 0.31% • Population in 2050: 226.3 million GDP: • Annual GDP growth rate (2015-50): 2.87% • GDP in 2050 (billion US$): 5370.1	Population: • Annual growth rate(2015-50): 0.31% • Population in 2050: 226.3 million GDP: • Annual GDP growth rate (2015-50): 2.75% • GDP in 2050 (billion US$): 5257.9	Population: • Annual growth rate(2015-50): 0.40% • Population in 2050: 231.9 million GDP: • Annual GDP growth rate (2015-50): 4.01% • GDP in 2050 (billion US$): 7042.9
Source	[24,25]		[26]

(c) COPPE

[i] More precisely, it should be the CO_2 equivalent of greenhouse gas emissions

[ii] This global carbon budget represents a high probability (above 0.66) to keep global warming levels below 2 °C by 2100[54]

[iii] The Brazilian Land-Use and Energy System model (BLUES) is an application of the MESSAGE platform. It is a mixed-integer linear optimization model, which minimizes the total cost of expanding the energy-land system to meet the expected demand for energy services and food. It combines technical, economic, and environmental variables for more than 8000 technologies with imposed constraints (including reforestation and deforestation scenarios) to obtain an optimal solution for the energy and Agriculture, Forest, and Other Land Use (AFOLU) sectors. The document is https://www.iamcdocumentation.eu/index.php/Reference_card_-_BLUES

[iv] The Computable Framework For Energy and the Environment (COFFEE) is a global and multi-sectoral partial equilibrium model with 18 regions (including Brazil) that runs on the MESSAGE platform. It uses the base year of 2010, with a horizon out to 2100 in 5-year time steps. The objective of the Model is to assess the potential synergies and trade-offs between energy systems and environmental and climate policy. The model includes all energy and land-use systems, with a hard link between the two. The macroeconomic inputs into the model come from exogenous macroeconomic drivers providing demand growth over time, or from the TEA model, which pulls from the SSP database.

Table 13. Comparative analysis of energy demand scenario studies on Brazilian

WEO2021	PNE2050	COPPE
Industry	Industry	Industry
Transport	Transport	Transportation
Buildings	Residential, Service	Residential, Commercial
Other	Agriculture, Non-energy use	Other Sector

(a) end-use sectors

WEO2021[i]	PNE2050	COPPE
Total	Total	Total
Electricity	Electricity	Electricity
Total liquids	Oil products, Diesel fuel, Other oil products, Fuel oil, Gasoline C, Hydrated ethanol, Kerosene/aviation gasoline	Liquids [ii]
Total gases	Natural gas, Liquefied petroleum gas (LPG)	Gases [iii]
Total solid fuels	Coal, Wood, Wood and charcoal, Sugarcane, Other	Solids

(b) energy carrier

[i]This is the aggregated end-use sectors used in this study
[ii]Includes ethanol, biodiesel and advanced fuels
[iii]Includes natural gas and LPG

Table 14. Correspondence between different studies

Criteria	Map	PV	Wind onshore	Wind offshore
inclusion	slope (°)	$m < 45°$	$m < 45°$	—
inclusion	distance to settlement (km)	$1 < m < 1000$	$1 < m < 1000$	—
inclusion	elevation (m)	$0 < m < 5000$	$m < 5000$	$-50 < m < 0$
inclusion	average wind speed (m/s)	—	0-50	0-50
inclusion	distance to coast (km)	—	—	$5 < m < 115$
inclusion	mining (0..1)	$m = 0$	$m = 0$	—
inclusion	salt/sand/ice (0..1)	$m = 0$	$m = 0$	—
exclusion	protected areas	$m \in \{1,\ldots,6\}$	$m \in \{1,\ldots,6\}$	$m \in \{1,\ldots,6\}$
exclusion	wetland	$m \in \{1,\ldots,10\}$	$m \in \{1,\ldots,10\}$	—

Table 15. Utilizable areas for the EnDAT analysis. m denotes the value of the according map to which the data is constrained, while provided integer categories are excluded.

Map	PV	Wind onshore	Wind offshore
bare	0.6	0.3	—
crops	0.24	0.15	—
grass	0.6	0.15	—
moss	0.6	0.3	—
shrub	0.6	0.15	—
forest	—	0.05	—
urban	0.024	—	—
marine water body	—	—	0.4

Table 16. Suitability factors for the EnDAT analysis. The land cover maps are given in shares from 0 to 1 and not mutually exclusive. Map data is taken from the Copernicus dataset[35]

.

Category	Parameter	Unit	Value
PV	power reduction	1/K	-0.005
PV	η_{module}	—	0.26
PV	η_{rest}	—	0.91
PV	availability	—	0.98
all wind onshore	nacelle height	m	112
all wind onshore	rotor diameter	m	165
all wind onshore	distance factor	—	6
all wind onshore	wind shading loss	—	0.85
all wind onshore	availability factor	—	0.982
wind onshore weak	nameplate capacity	kW	3630
wind onshore medium	nameplate capacity	kW	5330
wind onshore strong	nameplate capacity	kW	10550
wind offshore	nacelle height	m	150
wind offshore	rotor diameter	m	200
wind offshore	distance factor	—	6
wind offshore	wind shading loss	—	0.85
wind offshore	availability factor	—	0.95
wind offshore	nameplate capacity	kW	10000

Table 17. Technical parameters for the different generation technologies in EnDAT.

node_epsg4087.shp		
filed	***type***	***description***
name	string	Abbreviation of federal state
state_full	string	Full name of federal state in Portuguese
x	number	The latitude of the polygon center geometry of the federal state, CRS is EPSG:4087
y	number	The longitude of the polygon center geometry of the federal state, CRS is EPSG:4087

Table 18. Metadata of the records for 1 Defined region

EPEWebmap_equivalent_grid_aggregate_by_state.csv		
filed	***type***	***description***
node0	string	start node
node1	string	end node
transfer capacity	number	transfer capacity between the start and end nodes, in MW
efficiency	number	transmission efficiency between the start and end nodes assuming an efficiency of 1 for HVDC lines
name	string	The data processing produces a string that helps to trace each transmission line in the original dataset (EPE∼Webmap) by line name. The different line names are connected by the character "_".
length	number	length of the representative transmission between the start and end nodes
carrier	string	the type of the line, either AC or HVDC

Table 19. Metadata of the records for 2 Grid network topology

biomass_geographic_potential_reference_year_YYYY.csv		
filed	***type***	***description***
state	string	abbreviation of federal states
value	number	the installable capacity, in MW
reference_year	number	reference year, i.e., `YYYY`
type	string	power plant type – "biomass"
phase	string	Operational status. All values here are "potential" indicating the installable capacity, which is used to differentiate the status in the data of power plant infrastructure.

Table 20. Metadata of the records for 4 Installable capacity for biomass thermal plant

ANEEL_powerplants_per_state_per_type_reference_year_YYYY.csv		
filed	***type***	***description***
state	string	abbreviation of federal states
type	string	the type of power plants type – biomass, solar_pv, on_wind, mini_hydro, small_hydro, hydro, nuclear, coal, gas, oil
phase	string	operation status – operation or planning
value	number	capacity in MW
reference_year	number	reference year, i.e., `YYYY`

Table 21. Metadata of the records for 6 Power plants infrastructure

Hourly_sectoral_electricity_demand_per_state_YYYY.csv		
filed	***type***	***description***
time	string	the time stamp, `DD.MM.YYYY HH:00:00`
state	string	abbreviation of federal state
value	string	load value in MW. Note: As this is derived from the grid operator ONS, it includes the physical loss of SIN.
sector	string	end-use sector name – Buildings, Industrial, Other, Total (sum of each sub-sector)

Table 22. Metadata of the records for 7 Electricity load profiles

Energy_demand_scenarios_by_sector_by_energy_carrier.xlsx		
filed [xii]	***type***	***description***
Publication	string	the source of the data
Scenario	string	the full name of the scenario
Region	string	the name of the country, i.e., "Brazil"
Category	string	the indication of the data category. As it is the dataset of energy demand, it is "Energy".
Product	string	the energy carriers with aggregation. Values are "Total", "Electricity", "Total liquids", "Total gases", "Total solid fuels". Note: "Total" is the sum of the remaining energy carriers.
Flow	string	end-use sectors with aggregation. Values are "Total final consumption ", "Transport", "Buildings", "Industry", "Other". Note: "Total final consumption" is the sum of the remaining end-use sectors.
Unit	string	unit of the demand value, i.e., PJ.
Year	numeric	year. Values are "2010", "2015", "2030", "2040", "2050".
Value	numeric	value of the demand. The decimal point is written in ','.
Alias	string	The alias of the scenario used for plotting. It has the format `XXXX_YYYY`. `XXXX` is the abbreviation of the study, i.e. "WEO2021", "PNE2050", "COPPE". `YYYY` indicates the abbreviation of the scenario name, see table 13.

Table 23. Metadata of the records for 8 Scenarios of energy demand

Cross-border_transmission_RS-URU_RS-ARG_2012-2020_hourly.csv		
filed	***type***	***description***
time	string	the hourly time stamp, `YYYY-MM-DD HH:00:00`
node0	string	start node with Brazilian federal state abbreviation, namely, RS
node1	string	end nodes for neighboring country abbreviations, i.e., ARG and URU
power	number	electricity exchanged, MW

Table 24. Metadata of the records for 9 Cross-border electricity exchanges

Plant name	$\text{corr}_{\text{Pearson}}$
Fontes Solar I	0.92
Fontes Solar II	0.92
Assu 5	0.87
Conjunto Fotovoltaico Bom Jesus	0.86
Conjunto Fotovoltaico Ituverava	0.88
Conjunto Fotovoltaico Lapa	0.85
Conjunto Fotovoltaico Pirapora 2	0.83
Conjunto Fotovoltaico Nova Olinda	0.86
Conjunto Fotovoltaico BJL Solar	0.62
Conjunto Fotovoltaico Floresta	0.82
Conjunto Fotovoltaico Horizonte MP	0.76
Conjunto Fotovoltaico Guaimbe	0.78

Table 25. Correlation between PV site power generation and EnDAT simulation results for these sites

Plant name	$\text{corr}_{\text{Pearson}}$
Praia Formosa	0.44
Icaraizinho	0.51
Malhadinha 1	0.23
Alegria II	0.32
Alegria I	0.38
Elebras Cidreira 1	0.58
Xangri-LA	0.52

Table 26. Correlation of wind power generation between wind farm observations and EnDAT simulations