

Data Harmonisation for Energy System Analysis – Example of Multi-Model Experiments

Gardian, H.^{a,1}, Beck, J.-P.^b, Koch, M.^c, Kunze, R.^d, Muschner, C.^e, Hülk, L.^e, Bucksteeg, M.^f

^a Department of Energy Systems Analysis, Institute of Networked Energy Systems, German Aerospace Center (DLR), Curierstr. 4, 70563 Stuttgart, Germany

^b Helmut Schmidt University / University of the federal armed forces Hamburg, Institute of Automation Technology, Holstenhofweg 85, Hamburg 22043, Germany

^c Oeko-Institute e.V., Merzhauser Str. 173, 79100 Freiburg, Germany

^d Energy Systems Analysis Associates (ESA²), Bernhardstr. 92, 01187 Dresden, Germany

^e Reiner Lemoine Institut (RLI), Rudower Chaussee 12, 12489 Berlin

^f House of Energy Markets and Finance, Essen, Germany, University of Duisburg-Essen, Universitätsstr. 12, 45117 Essen, Germany

Abstract

A variety of models have emerged in the field of energy system analysis to answer a wide range of research questions surrounding a sustainable future for the energy sector. Even models that are designed to address similar issues often have a different focus or modelling approach. Thus, model experiments are an important tool to provide an overview of the range of models and to enable decision makers to make meaningful model choices. The execution of such comparisons is based on a harmonised data set to ensure a high degree of comparability. In the MODEX project cluster, six model experiments including 40 energy system models were conducted and efforts were made to harmonise the input data not only within the individual comparisons, but also beyond them in the consortium. The experiences and findings of the consortium on how data harmonisation could be performed are presented in this paper. In particular, the focus lies on data transparency to ensure a high degree of reproducibility. A key finding is that while model heterogeneity complicates harmonisation, an early focus on data research and scenario design promotes the creation of a common data set. The collection of metadata can provide a significant advantage for the use of model experiment results by external scientists but also for the data acquisition process itself because of the predefined machine-readable and standardised format.

Highlights

- Experiences in data harmonisation at the example of 40 energy system models
- Identification of potential problem areas in connection with data unification
- Introducing methods to increase traceability and transparency in model comparisons
- General findings and suggestions for the execution of data harmonisation

Keywords: Data harmonisation, energy systems analysis, model comparison, metadata, energy systems modelling,

¹ Corresponding author

Email address: Hedda.gardian@dlr.de (Hedda Gardian)

Word count: 7141

List of abbreviations

BDI	Federation of German Industries
DWH	Data Warehouse
EMF	Energy Modeling Forum
EMS	Energy Models System
ENTSO-E	European Network of Transmission System Operators for Electricity
GDP	gross domestic product
NUTS	Nomenclature of Territorial Units for Statistics
OEP	Open Energy Platform
RE	renewable energy
TYNDP	Ten-Year Network Development Plan
UNFCCC	United Nations Framework Convention on Climate Change

1. Introduction

1.1. Background and motivation

In recent decades, due to the pressing issue of climate change and the resulting energy transition, a large number of models for energy system analysis have become established, which differ from each other in their methodological approach and focus [1]. These energy modelling activities influence the development of energy systems and are therefore of public interest and require accountability and transparency. While in the past, modelling activities have remained rather opaque, the energy research community is expanding to facilitate reproducibility in methods and to share research processes more openly [2]. With this recent ever-increasing demand for more transparency, the topic of open input data, open software and accessible result data is coming more into focus. Efforts have already been made to make the models available to a broad community and thus to increase the exchange between modelers [3]. However, the field of systems analysis remains convoluted for decision makers as to which model choice to make for a specific research question. One way to gain more insight into the increasingly complex models is to conduct methodological model comparisons. On the one hand, this aims to increase transparency in the model landscape for outsiders, such as project planners or policy makers. On the other hand, it also provides scientists the opportunity to improve their models or validate their model results. Modelling hurdles can be overcome, as the methodology of model experiments enables to increase the understanding of one's own model and to identify the strengths and weaknesses of the respective modelling approach.

Another key to transparency in energy system analysis, beyond the mere provision of data, is the precise description and origin of these by means of metadata using the FAIR principle [4]. If all data used in model experiments are publicly available and sufficiently described, as well as the procedure for comparisons documented, a model experiment can be a great enrichment for the modelling community, because only in this way traceability is guaranteed.

One approach to these principles is the MODEX model experiment cluster. It consists of six research projects (FlexMex, MEO, MODEX-EnSAves, MODEX-NET, MODEX-POLINS and open_MODEX) in which in total 39 partners with 40 models perform model comparisons on current issues in systems analysis [5]. Each project has an individual thematic focus within the broad spectrum of energy system analysis topics. Accordingly, they differ in the set-up of the involved models and the spatial, temporal, and technical resolution of the underlying scenario (see Figure 1).

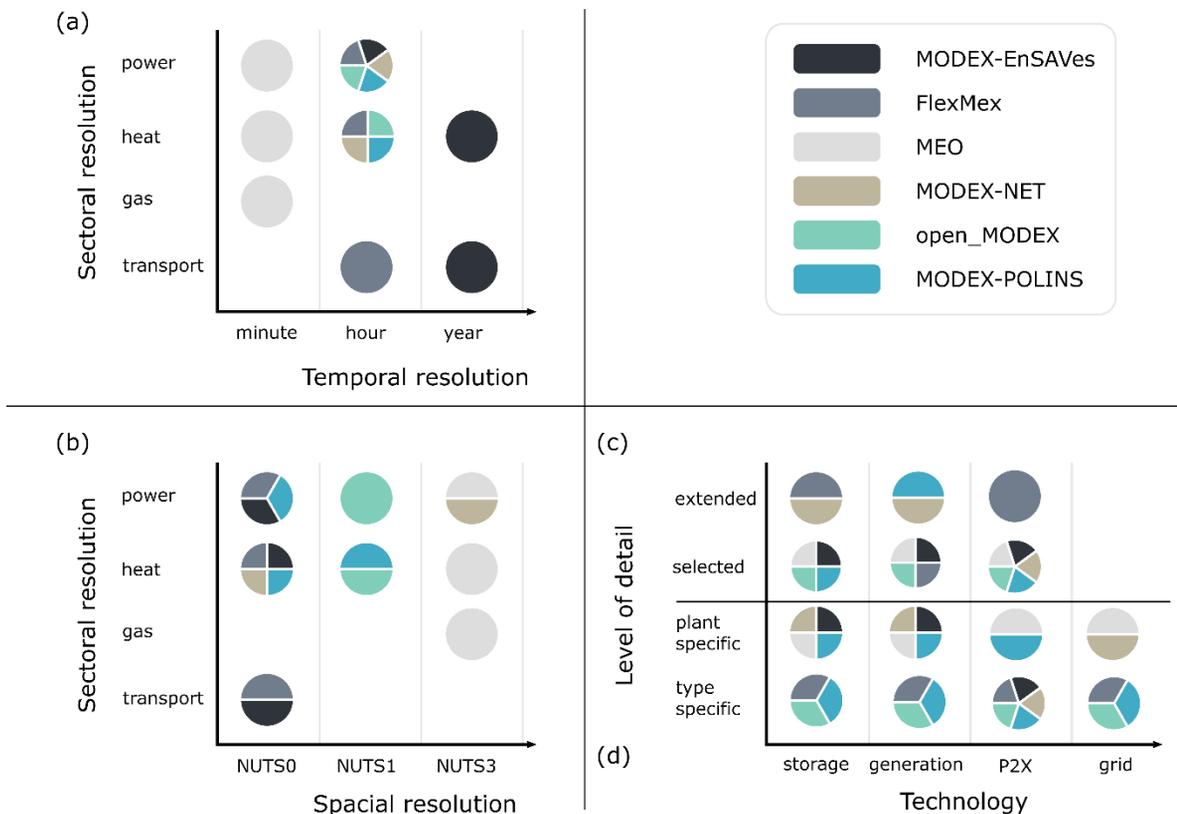


Figure 1: Classification of MODEX projects. Figure (a) shows the temporal resolution of the individual sectors, while figure (b) depicts the regional resolution. The two-part figure regarding the level of detail in the lower right quadrant shows at the top (c) whether a large range of technologies of a class or only selected ones were analysed and in the lower area (d) whether plants were modelled aggregated or individually. If circle segments of the same colour appear twice in a graphic, the individual models within a project fall into different categories.

Although both projects, MODEX-EnSAves and FlexMex, deal with sector coupling technologies, they differ significantly in their objectives. While FlexMex has a strong technological focus with a comparison of various load balancing options, MODEX-EnSAves analyses the influences of the development of electromobility and power-to-heat technologies on security of supply. In MEO, there is a methodological emphasis on comparing the system operation, whereas in MODEX-NET, models with a similar technical focus on network modelling are compared. The open_MODEX project takes a broader approach, focusing on frameworks with open licenses. The model comparison in MODEX-POLINS includes models for the evaluation of policy instruments, such as carbon pricing or coal phase-out plans.

Both the selection of a suitable modelling methodology and the selection of adequate input data have a great influence on the quality and significance of the results of a model-based system analysis. In the MODEX cluster, methodical model comparisons are in the foreground. To increase comparability across the modelling experiments, the input data of the models are harmonised as far as possible.

1.2. State of research

The history of model comparisons in energy system analysis goes back several decades. Accordingly, the Energy Modeling Forum (EMF) at Stanford University was established as early as 1976. However, the focus of past model intercomparisons was less on open-source models, open-source data, or transparency and traceability, but rather on defined thematic areas. Some examples of model comparisons performed in the past are listed in the following Table 1 in chronological order.

Table 1: Literature review of conducted model experiments.

Model experiment	Reference	Focus
Forums or platforms for performing model comparisons		
EMF	[6]	With currently 37 ongoing or completed studies the EMF is one of the most extensive platforms for model comparisons. The focus lies on answering energy and environment related research questions, such as the investigation of strategies to mitigate climate change. Thereby, a number of different models are used in order to examine the topic from various perspectives; a validation and the investigation of causes for model deviations play a minor role.
FORUM	[7]	A collaboration of several model experiments took place in the early 2000s within the Forum for “Energy Models and Energy Economic Systems Analysis” (FEES). The aim was to analyze the relevance of alternative methodological approaches as well as different levels of detail and data assumptions on the results.
EMP-E	[8]	Launched in 2017, the “Energy Modelling Platform for Europe” (EMP-E) is providing another platform for the exchange of data, methods and developments.
Individual model comparison projects		
CASCADE MINTS	[9]	In the European project “Case Study Comparisons And Development of Energy Models for INtegrated Technology Systems” (CASCADE-MINTS), electricity coupled with hydrogen scenarios were investigated in two parts. First, the models involved were extended to represent the relevant technologies and then the influence of these key technologies and policy options were jointly investigated.

ENSAMBLES	[10]	In order to obtain more reliable and accurate results for the prediction of climatic developments, future scenarios were calculated in the Ensembles project with the aid of ten global climate and earth system models, on the basis of which probabilistic estimates of uncertainties for the future climate could subsequently be carried out.
ADVANCE	[11, 12]	In another European project called ADVANCE, IAMs were compared and used to investigate the implications of the Paris Agreement.
RegMex	[13, 14]	A model experiment with the aim of publishing and documenting detailed data sets was RegMex, which aimed to derive robust conclusions on the transformation of the energy system.
BEAM-ME	[15]	The BEAM-ME project has a distinctly different emphasis, by exploring the potential for highly parallelised computing to accelerate complex energy system model calculations and to optimise them for supercomputers. The results were applied to energy system models in a subsequent model experiment.
4NEMO	[16]	The 4NEMO project aims at integrating economic as well as social dynamics and their related uncertainties into energy system models. The results obtained can indicate which are the strengths, weaknesses and advantages compared to other models
AMPERE	[17]	In the international “Assessment of Climate Change Mitigation Pathways and Evaluation of the Robustness of Mitigation Cost Estimates“ (AMPERE) project, 17 energy-economy and integrated assessment models are participating to map the possible pathways to meeting climate targets while taking climate policy into account. Here, too, model comparisons are used to determine the uncertainty of the results and, based on this, differences are examined to improve the model-based analysis.

Content-related focal points play only a subordinate role in the MODEX cluster. Instead, particular emphasis is placed on methodological comparisons of the models, open data sets, and a high degree of transparency and reproducibility. Of the 40 participating models, 12 feature an open code base and are openly documented. These include Balmorel, DIETER, Energie-Agent (partly open), eTraGo, GENESYS-2, GENESYS-MOD, IntegraNet/Transient, mosaik, μ GRIDS, oemof, pandapower and Urbs.

The topic of data preparation and data harmonisation for energy system analyses has not been focussed on in the research landscape so far but is becoming increasingly important. Still, no uniform standards for semantics, structuring and storage of energy system data have been established. One of the main reasons for this is probably the fact that the methodological development of models and the answering of research questions are often the main focus of system analysis. Since the models are usually developed as stand-alone applications, their data management is often also organised individually and tailored specifically to each individual model. This complicates the transferability and thus the reusability of elaborately created data bases in the research landscape of energy system analyses.

Even when relevant data for energy systems analysis is made publicly available, there is often a lack of appropriate licensing, lack of indexing to find the data or missing contextual information to evaluate the data factually [2].

Evolving metadata strings to describe published data attempt to address these shortcomings, but are themselves very heterogeneous, due to the wide variety of energy data to be described [18, 19]. To ensure a minimum of openness and reusability through metadata, FAIR principles and the principles of open science have been integrated into the development of metadata, which has led to some standardisation [4, 20]. In this work we use a metadata string that follows these standards to describe the data and increase the transparency from the different MODEX projects.

1.3. Contribution of this paper

The paper focuses on methods for data harmonisation and data documentation as well as on the topic of open data, which has been given too low a priority in energy system analyses so far, but that is indispensable for their transparency. In addition to methodological model comparisons, the MODEX cluster pays special emphasis to these issues. The aim is to stimulate the exchange of experience between modelers on these topics and to identify barriers and basic approaches to handle it. With this in mind, not only the necessary data matching for the implementation of the methodical model comparisons within the individual MODEX projects will be carried out, but also an effort will be made to achieve a high degree of harmonisation across the projects. This is intended to raise modelers' awareness of these important issues and to provide the community with the opportunity to benchmark their own model by means of detailed documentation of used data and methods.

The cross-project data coordination was the responsibility of the group of data managers, which consisted of one partner from each of the MODEX projects. Sections 2.1 to 2.3 describe the approach and results of the data managers group in the MODEX cluster. The focus was on the most important framework data for modelling (esp. wholesale fuel prices, emission factors, etc.), which are usually required by a large part of the models used in the entire network and at the same time represent essential model drivers.

The cross-project data management formed the basis for the necessary further data harmonisation within the individual MODEX projects for the model comparisons. Depending on the analysis focus of the models used, different approaches were followed. Section 2.4 presents the methodological approach of MODEX-EnSAves as a representative example, since in this project not only models are compared, but also coupled with each other through a common data base. The latter places special demands on data harmonisation, which are illuminated against the background of this still rather young trend in the field of energy system analyses [21]. Section 2.5 is then dedicated to the topic of open data and the documentation and reusability of model data. Finally, chapter 3 collects, clusters, and describes the hurdles and challenges encountered during the harmonisation process. Approaches to solutions and resulting consequences are explained accordingly.

2. Materials and Methods

In the following, we will describe our data harmonisation procedure in more detail, focusing on the harmonisation of input data. For these, several harmonised modelling aspects could be identified.

2.1. General modelling aspects

First of all, general modelling aspects were examined. These were divided into two groups. Under the spatial system boundaries group, the regions & countries, Nomenclature of Territorial Units for Statistics (NUTS) levels and network levels of the MODEX projects were compared. Under the analysis period group, the start year, analysis horizon, base year, temporal resolution, weather year, and real price reference year were collected. For the analysis period, several aspects could be harmonised, while others remained unequal due to the different focal points of the studies.

Regarding the considered regions & countries, MODEX-EnSAves, open_MODEX and MEO perform their investigations only for Germany or single subregions. MODEX-POLINS, FlexMex and MODEX-NET, however, take the neighbouring and other European countries up to all European Network of Transmission System Operators for Electricity (ENTSO-E) countries into account. In terms of regional resolution, some

projects consider details at NUTS 3 level, while the majority distinguish only NUTS 0 regions. Different grid levels are considered in MEO only, where the low-voltage grid is modelled as well.

For the analysis period, the start year was harmonised for all six MODEX projects such that all use 2016. The analysis horizon varies significantly from project to project. While MODEX-POLINS, MODEX-EnSAves and MODEX-NET consider an analysis period up to the year 2030, FlexMex and open_MODEX chose the year 2050. The MEO project does not perform a classical system analysis and thus has no analysis horizon, but relates its investigations exclusively to the start year. The base years vary, of course, due to the different analysis horizons. While MODEX-POLINS chose 2025 and 2030 as base years, FlexMex uses only 2050 and open_MODEX considers base years in intervals of 10 years. Without an analysis horizon, MEO strictly speaking has no base years, but forecasts for 2024 and 2034 are used. MODEX-EnSAves performs calculations for 2020, 2025 and 2030 and the MODEX-NET project chose 2016 and 2030 as base years. The temporal resolution is very similar in almost all projects, as calculations in the electricity sector are carried out on an hourly basis. Only MEO has a higher temporal resolution of 15 minutes, for its detailed network calculations. For the weather year, all projects except FlexMex (2012) and MEO (2011) chose the start year 2016. MODEX-EnSAves and MODEX-NET use additionally other weather years, e.g. to simulate an extreme weather year. For the price reference, all projects uniformly use 2016.

2.2. *Gathering of the data requirements*

While the focus was previously placed on the harmonisation of general modelling aspects, the harmonisation of input data will now be considered in the following.

Depending on the type and focus of the individual models, there are different requirements for the input data needed, for example with regard to their spatial and temporal resolution, but also generally for individual categories of input data. For instance, power grid models as applied in MODEX-NET and MEO inherently rely on input data on the topology and the degree of expansion of the power grid and also require all data on power generation and demand in a corresponding spatial resolution at the level of the grid nodes.

In order to get an overview of the required input data and also of the resulting outputs of the 40 models used in the individual MODEX projects, these were systematically recorded in an input-output table on the basis of various categories. The eight overarching data categories are:

- Macro-economic and statistical data, including subcategories for gross domestic product (GDP), population, employees, households, buildings, freight volume, policy objectives, etc.
- Environmental data, including subcategories for weather data, fuel type and technology specific emission factors and land usage, etc.
- Demand data, including subcategories for demand volumes and profiles for electricity, heat, fuel types and other energy carrier as well as for CO₂ needed in methanation, etc.
- Installed infrastructure data, including subcategories for power plant portfolios, storages for electricity, heat or gas, electricity, heat and gas grids, vehicle fleets, load points for electric vehicles, etc.
- Technology- and plant-specific parameters, including subcategories for efficiencies, load gradients, specific investments, lifetimes, specific costs, availabilities, etc.
- Prices and costs, including subcategories for fuel, electricity, heat, CO₂ for methanation, CO₂ certificate and redispatch costs, etc.
- Actor behaviour and acceptance, including subcategories for self-consumption maximisation, driving profiles, remediation activity, etc.

- Deployment/utilisation of infrastructure, including subcategories for deployment profiles for generation and storage of electricity and heat, utilisation of electricity grids, congestion management, expansion or deconstruction of infrastructure, etc.

As expected, the input-output table result shows a wide range of input data used across the more than 150 subcategories. In addition, some data are input data for some of the models, while they are part of the outputs for others. For example, this is the case for dispatch-only models compared to investment models for installed generation capacity. The input-output table is online available at [Link]².

Thus, only a small subset of the possible input data can be considered for data harmonisation, which enters all models as input data and for which data harmonisation also appears to make sense. The following input data were identified as the lowest common denominator across all models and projects: Prices for fuels and CO₂ certificates, fuel-specific CO₂ emission factors, country-specific load profiles for electricity and district heating, and exchange capacities (NTC) at the interconnection points in the European electricity grid.

In order to ensure the highest possible transparency regarding the input data used in the individual MODEX projects in addition to the data harmonisation, their metadata were collected and published (see 2.5).

2.3. Description of model input data

As part of the data harmonisation process, a literature search was carried out for the individual input data and a default value was formed for each. For the development of fuel and CO₂ prices, the "Current policies" scenario from the World Energy Outlook 2018 was selected as the base source [22]. In addition to the prices for CO₂ certificates, this affects the energy carrier hard coal, natural gas and oil. For lignite, the Grid Development Plan for Electricity 2030 (version 2019) was used [23], for nuclear energy the Ten-Year Network Development Plan (TYNDP) 2018 [24] and for biomass products the assumptions of the reference scenario from the Federation of German Industries (BDI) study "Climate Paths for Germany" [25]. All prices were converted according to the price reference year 2016.

The fuel-specific CO₂ emission factors were determined in accordance with the United Nations Framework Convention on Climate Change (UNFCCC) reporting and based on the national inventory report on the German greenhouse inventory [26]. For lignite, the average value between East German and West German coalfields is calculated and for biomass the CO₂ emission factors are set at zero.

The load and renewable energy (RE) profiles for 2016 are based on the ENTSO-E Transparency Platform and have been normalised to values between 0 and 1 in terms of installed capacity or maximum electricity demand [27]. Based on these normalised profiles, the absolute profiles can then be formed as a function of the annual energy quantity for the individual scenarios.

The input data harmonised in this way was made available to all MODEX projects as default values via csv files. If possible, these default values should be used in the project-specific scenarios. If this was not possible, other data could also be used. This was then briefly justified in the respective scenario documentation.

² The files are currently being prepared for publication on the Open Energy Platform. For the review, they are provided as supplementary material. Please note that classification of the frameworks only refers to the requirements of the scenarios in the specific projects. The frameworks are more versatile.

Table 2 shows an overview of the proposed data sets and to what extent a harmonisation could be conducted, where individual projects deviate from the suggested data, or even whether the proposed data set is not used in any project.

Table 2: Data harmonisation – model model inputs of scenario dependent parameters.

Parameter	Suggested source	MODEX-POLINS	FlexMex	open_MODEX	MEO	MODEX-EnSAves	MODEX-NET
Electric load profile	Eurostat 2018/a	X	X + ENTSO-E 2016	Own Data	SimBench	X + me	X
Heat load profile	Bründlinger et al 2018, IEA 2018a, Renewables.ninja 2020	X	Gils2015	Own Data	Own data	X	Own Data
RE generation/ RE feed-in	OPSD 2020, Entso-E 2018b	X	Own data from EnDat-Modell, Scholz2012	me	me	X	X
Emission factors/ energy generation	UBA 2016 [26]	(X)	X	X	X	(X)	X
Installed generation capacity	Bundesnetzagentur 2019, Entso-E 2018b	X	me	MaStr2021	me	X	X
NTCs	Entso-E 2018b, Rippel et al 2019	X	X	DLR SciGrid	X	X	X
Fuel prices	IEA 2018b, BMWi Langfristszenarien, da Szenariojahr 2050	X	X	X	X	X	X
CO ₂ price	"IEA 2018b, Hirst and Keep 2018, BMWi Langfristszenarien, da Szenariojahr 2050"	X	X	Own data	X	WEO2018 (Scenario "Sustainable devel.")	X
	X	harmonises with the proposed MODEX values					
	(X)	almost harmonised with the proposed MODEX values, only slight deviation					
	me	model endogen					
		other data set / own data set					

Due to the very different research questions of the individual MODEX projects, a complete harmonisation was not possible and this would also not have been expedient. However, Table 2 shows that some parameters could be harmonised across the projects. The first column describes which data is involved in each case, e.g. the electricity load profile. In the second column, the proposed data set or the source is mentioned. The following columns list for each MODEX project whether the proposed source has been used or if there are deviations from it. These deviations can be minor, in which case it is considered as harmonised. However, data values can also be generated endogenously in the models and thus are not included in the models as a data set at all. In addition, the table indicates as well the utilisation of other data sources, these are then labelled with the respective source.

In summary, it can be said that the emission factors, the prices for fuels and CO₂, and the NTCs have been harmonised well. For the electricity and heat profiles, it is clear that harmonisation was more difficult, as can be seen from the use of various data. The same applies to the RE generation and the installed electrical capacity.

2.4. Need and Background of data harmonisation for model coupling – example of MODEX-EnSAves

The challenge of harmonising models, which all perform the same task, becomes apparent in the described methodology and the limited number of harmonised data sets. An additional component of complexity arises when these models do not compute in parallel but in cascade.

Many model developments take a comprehensive system view and depict the relevant stages of the energy supply chains in a highly aggregated or simplified manner, but as completely as possible. In recent years, however, there has been a trend for many model developments to focus only on individual aspects or sub-areas of the energy system (e.g. on the development of energy demand in the transport sector or on the design of regional energy markets), but to model these in much greater detail. [21]

This is essentially due to two factors. Firstly, during the energy transition, an enormous variety of new stakeholders and options for action in system design are emerging. On the other hand, the requirements from politics and industry for the findings of system analysis have increased significantly. In particular, more robust assessments of economic viability, market success or investor behaviour as well as the consideration of acceptance in the economy and the society are becoming more important. At the same time, the holistic assessment of the impacts of alternatives to system design in terms of life cycle assessment or with regard to social aspects is becoming increasingly important.

This development has led to much more detailed and thus more complex research questions, which usually cannot be modelled and analysed in the required depth of detail with a single comprehensive system model. Nevertheless, the cross-sectoral system view still plays an important role in order not to neglect the existing interdependencies between the individual parts of the energy system. In order to be able to adequately take these interactions into account, the specific models for individual system components must be coupled with each other and, if necessary, also with models with a comprehensive aggregated system view and to form an Energy Models System (EMS) [28]. The coupling primarily concerns the data exchange between the models, in which the results of one model serve as input data for other models. The model application within the EMS usually takes the form of an iterative process. This process may require several iterations to converge the results of the models involved.

Due to the numerous degrees of freedom in modelling as well as the lack of standards for structuring and holding the necessary data for the energy system analyses, this point represents the greatest challenge in model coupling. The creation of an EMS is usually done in the following three steps:

Step 1: Selection of the models to be used and definition of the interfaces for data exchange

The selection of the models to be used essentially depends on the problem to be investigated, the boundaries of the system under consideration, and the required spatial, temporal and technical resolution of the modelling. The model selection must be made in such a way that all relevant system aspects and their interactions are covered with the necessary level of detail.

To define the interfaces for data exchange, the defined input data are first compared with the output data of all models involved and the overlaps of input and output data sets for the possible data exchange are identified. It is recommended to focus on the most important model drivers. On this basis, the needed interaction of the individual models within the EMS is designed and agreed upon. It must be taken into account that the models used may well have overlaps in the modelling of individual system aspects. For this reason, in addition to determining the data sets to be exchanged, the modelling of the individual system aspects must also be clearly delineated and the model responsible for each must be determined. At the same time, the handling of necessary data transformations during data exchange between the models is methodically determined. This can concern, for example, the breakdown of spatially aggregated results to local type regions or the conversion of wholesale prices into end-user prices.

Furthermore, a detailed model application plan is created in this step. This contains the sequence and the schedule of the individual model applications and schedules the respective data exchange. The model application plan is particularly important if models from different institutions are to interact smoothly, as delays in one model usually affect all models involved due to the iterative model application process.

Step 2: Generation of a common harmonised database

The second step essentially corresponds to the procedure of the data harmonisation in the MODEX cluster already described in section 2.2. From the matching of the data sets in step 1, the input data sets of all participating models with the same content are identified, which are not provided as an output by one of the participating models within the framework of the data exchange, i.e. which do not represent endogenous variables within the EMS. These are essential model drivers that usually represent exogenous influencing factors for the development of an energy system (e.g. the development of the population or industrial production) or are to be defined at the defined system boundaries (e.g. world market prices for energy carriers). This information forms the common framework data that must be harmonised between the models in order to ensure a consistent analysis within the EMS. Special attention should be paid to how a specific data set is used in the individual models. For example, in some models interest rates are used to model specific stakeholder preferences, while in others they represent the usual discount rate in an economy or industry.

Step 3: Configuration of the model interfaces for data exchange

The third step involves defining the required data mapping. Since the models to be coupled were usually developed independently of each other as stand-alone applications with separate data storage, they usually have significant differences in the data set structures. As a further harmonisation aspect, the mapping thus forms the basis for achieving data compatibility between the models and is therefore also a decisive prerequisite for their consistent interaction. Challenges arising here are of various types:

- The models use different levels of detail or aggregation of the data (e.g. regarding countries, sectors, technologies, energy carrier, etc.).
- Different identifiers for same datasets exist (count and labelling) and there are different restrictions for the labelling of identifiers in the models (allowed characters and length).
- The time structures of same datasets can differ (yearly, seasonal, hourly, etc.).
- The provided or required data have different file formats (e.g. csv, sql, etc.).

Another major challenge arises from the size of the data sets to be exchanged, which results from the increasing requirements for high spatial and temporal resolutions in system modelling. A small example to illustrate the

dimensions that now occur: A single data set on German electricity demand, broken down to NUTS 3 level for 8760 hours per year, already reaches a size of over 10.5 million data points with an analysis horizon of 30 years with only 3 modelled interpolation years. If the electricity demand was further differentiated into 50 individual energy application processes, e.g. for adequate modelling of load management measures, more than 500 million data points would be reached. Applied to an EU-wide modelling, this data set would comprise over 1.7 billion data points. These dimensions place enormous demands on the common model data management and the tools for data exchange. Specified models are necessary to handle these challenges efficiently.

These data management tools include functionalities such as the integration of harmonised data (step 2) into a common database, the provision of an interface for data exchange (step 3) and the linking of input data to result data for publication, ideally combined with the corresponding metadata. There are some tools that already support these or similar functionalities, such as the data warehouse (DWH) of the project partner ESA² used in MODEX-EnSAves [29], but also pyam [30], Spine [31] and the Open Energy Platform (OEP) [32] in combination with the oedatamodel [33] can automate the coupling process. The utilisation of those tools makes it possible to meet the increasing requirements for transparency of system analysis as well as documentation and reusability of data. These essential aspects are considered in more detail below.

2.5. Required metadata management

As a consequence of the methods and experiences described, sound data and model management is necessary for modelling exercises, but they also must be transparent to ensure good scientific practice. Metadata, as a flexible vehicle to document information of information, supports the development towards more transparency along the modelling process from scenario and model definition, selection of input data, pre- and post-processing of data to results communication. Of these steps, the documentation of input data is the most cost-effective and promising to increase transparency due to its low complexity and evolving documentation standards within subdomains of energy modelling [18, 19]. The other process steps' documentation and comparability across different models and modelling exercises remains more complex or less uniform (in the domain of energy modelling), despite standardisation efforts by various research groups. These efforts rely largely on metadata and focus on scenario documentation and interoperability; the linkage of existing data bases and provenance documentation on modified data sets; the reproducibility and reliability of scenario processing; and the development of an open energy ontology as accelerator for improved interoperability [33, 34, 35, 36, 37]. These projects follow principles of open science as guidelines to promote transparency and reproducibility. Some MODEX modelling experiments publish their input data, allowing reproducibility, and some also under an open license for free re-use.

This reflects the fact that transparency and reproducibility in energy system research is still not a matter of course, despite the recent positive developments [3]. Although documenting and publishing the data is the simplest form of contributing to transparency, it is not yet a complete contribution to its reusability. Data sources must be licensed and documented in the metadata accordingly (the same applies for published methods) [38]. Enough meta-information must be given in the metadata to ensure clear data interpretation. Data sets in energy modelling, however, are very large and the types of data are very heterogeneous, stemming from fields of geography, meteorology, economics and engineering, posing challenges to metadata documentation. In particular, the requirements for the level of detail of documentation differs between domain experts, who tend to need a higher density of information, and the non-domain experts, whose information needs are less detailed in order to ensure subsequent use of the data. Metadata documentation must be flexible enough to meet these needs to form a standard in the domain of energy system modelling.

Despite the incompleteness of data publications in the MODEX cluster, at least metadata on selected modelling parameters is compiled [Link]³. Challenges remain in publishing open datasets, often due to a lack of open and properly licensed data. When republishing data openly, the rights of the copyright holders of each data point must be known in order to choose an appropriate open license. Furthermore, the chosen license should be as unrestricted as possible to be compatible with other open licenses and to avoid data silos. Funding institutions such as the European Union have recognised this and increasingly demand the use of open data in tenders. The minimum target of publishing metadata for the input data has been defined in the MODEX cluster to increase transparency in research and to take a step towards reproducibility. Ultimately, the collected metadata will facilitate understanding and comparability of the modelling activities, increase transparency of scientific practice, promote reproducibility of research and thus lead to better interpretation of the research output. The metadata evaluations of the MODEX cluster show though, that the input data across the projects is largely under copyright protection.

Among the MODEX model comparisons the greatest possible data harmonisation among selected input parameters has been targeted. However, as seen above, this has only been achieved partially due to the diversity of participating energy models or energy modelling frameworks and their largely heterogenous need for data input and scenario frames and model assumptions. Its documentation across all projects, the has been facilitated with the *oemetadata* (v.1.4.1), as both human-and machine-readable standard of data sources, licenses and scenario assumptions [18]. The *oemetadata* string has been used in the energy system modelling community for a few years and in addition to other types of documentation it helps to find and to process data sources and scenarios more easily. It allows to document multiple resources for the parameterisation of an input parameters and provides detailed documentation possibilities of temporal and geographic information, licenses and more in-depth information. Simultaneously, the *oemetadata* follow the FAIR Guiding Principles for scientific data management and stewardship [39]. The FAIR principles promote guidelines for data stewardship established to facilitate the process of discovery, evaluation and reuse of data and other digital assets in publicly funded projects. The principles explicitly include “computational stakeholders” who, in addition to human reuse of resources, play an increasing role in exploring and processing data in data-driven research projects. The principles in short state, that a digital asset should be findable and described with rich metadata; it should be accessible and metadata accessible, even when the digital asset is no longer available; it should use interoperable formal and accessible metadata written in a broadly applicable language; and it should ensure reusability by indicating a clear and accessible license. The *oemetadata* (v.1.4.1) implement those principles and remain flexible in its structure to meet the future challenges of data interoperability.

3. Results and discussion

The difficulty of the targeted undertaking is the harmonisation beyond the individual model comparison projects in the MODEX cluster in order to establish comparability here as well. The different model and project foci reduce the possible extent of the unified input data. Nevertheless, using the methodology described above, an attempt was made to standardise the input data in the assumption that statements made as a result of one of the projects might be confirmed in other projects. Confirmation or disproval of this thesis is still pending. However, experience has already been gathered on the challenges that have arisen as a result of this approach. A survey among the participating modelers was able to capture the different aspects of the hurdles related to a

³ The files are currently being prepared for publication as a JSON-File on the Open Energy Platform. For the review, they are provided as supplementary material.

harmonisation of input data and to collect the resulting issues (see Figure 2). Participating modelers were asked what model-independent and model-dependent hurdles they encountered during the harmonisation process, both within and across projects. We identified four categories of problem areas: issues due to model differences, challenges related to the scenario definition, difficulties in adjusting data in the models, and in the data collection itself.

Due to the different level of detail concerning spatial, temporal and technological resolution, a reduction of the scenario definition and the required data respectively to the lowest common denominator is inevitable. Even though data requirements of the models within the projects show a significantly larger intersection than in the MODEX consortium, nevertheless, the consensus on a uniform data basis has led to scenarios of low complexity that deviate from common plausibility in some projects. While some projects managed to create meaningful case studies and results from an academic perspective, this goal was secondary for other projects, since the focus was placed on the comparison of the models and the scenarios were merely seen as a test case.

In addition, transferring common input data into the models' parameterisation can be error-prone and introduces hurdles. Due to the vast number of parameters that models must adjust, multiple iterations are required to correct errors in the transfer or to add forgotten data. This iterative process can be lengthy for model comparisons involving a cascade, that is, using model coupling as described in section 2.4. Also, adjusting parameters may be more time-consuming for models that use databases in the background. It has been shown that an early comparison not only of the results, but also of the input data that the models ideally provide when they report their results, simplifies the outlined process, shortens it, and makes it more manageable.

Another issue that can complicate the transparent comparison of models is the acquisition of the data itself. Since this project aims at maximum transparency and traceability, the use of open and licensed data sets is desired. The thus limited data pool often does not include scenario data and for the non-electricity sectors the data situation is scarce. As a result, the projection of current data and the use of expert guesses is often resorted to.

In order to quantify the impact of these problem areas, the survey also asked about deficiencies in the preparation and harmonisation of data and about necessary simplifications. The modelers have pointed out that a 100% harmonisation was infeasible. It was noted that errors in the parameterisation could not be excluded completely due to the possibility of a different understanding and use of parameters. Also, in some places, a direct conversion into the required unit is not possible (e.g., specification of transport service demand in person- or ton kilometers vs. specification of number and demand of individual passenger cars). Also, far-reaching simplifications have to be accepted in order to reach a common data basis, such as the aggregation to the same level of detail or the omission of parameters that would allow a more detailed description of technologies.

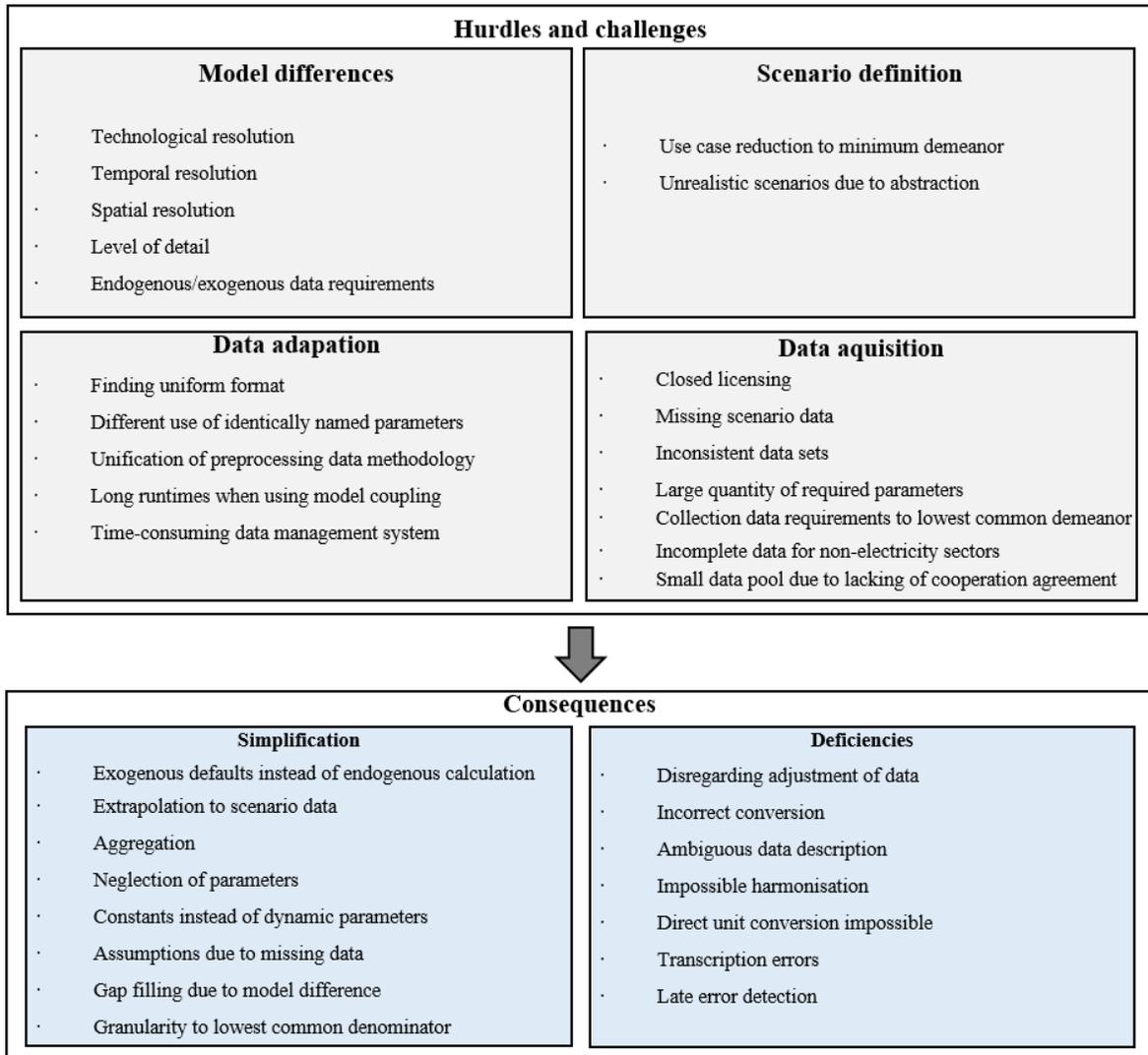


Figure 2: Collection of hurdles and challenges encountered during the harmonisation process and the resulting simplifications and deficiencies.

4. Conclusions

The quality of model comparison results is directly dependent on the degree of harmonisation of the input data. Only through a maximally harmonised data basis a distortion of the results due to deviating input data can be excluded. However, in addition to a harmonised database, which is decisive for the outcome of a model comparison, its documentation and licensing is essential in order to make the results usable for scientists beyond the model experiment. The experiences during the implementation of this process within the MODEX project cluster were described in detail and furthermore the applied metadata string was introduced. Nevertheless, the

work has shown that the procedure of data harmonisation challenges modelers with a considerable task. To point out the additional challenges of data harmonisation in a model coupling context, this topic has been taken up in a short excursus.

It could be observed that the described hurdles and issues concerning the harmonisation of input data for model comparisons are amplified according to how much the models under consideration differ from each other. Moreover, in the case of the MODEX cluster, the variety of project related research questions increased harmonisation challenges. Providing a common database was already a challenge within the individual MODEX projects themselves and a complete harmonisation within the MODEX cluster has not been possible. Only the unification of basic parameters, such as emission factors, was feasible to realise. The meaningfulness of a comparison of the model results beyond the project boundaries is therefore to be questioned. However, for the harmonised data a duplication of research work could be avoided.

Although, overarching harmonisation was not possible, the efforts of the group of data managers had several positive outcomes. By means of the overarching work, a rough framework could be created and consistency within the project cluster could be established. MODEX was a unique opportunity to develop a sound methodology for harmonisation in model comparisons, as it is rare to find such a large number of models in one project cluster, representing a broad range of the energy system analysis. With the help of the knowledge gained, it was possible to show how complex and time-consuming data harmonisation processes could be performed. In the process, several hurdles could be identified and solutions could be proposed.

Furthermore, several ways could be identified to improve a harmonisation process of input data for future projects. Among them is a detailed assessment of data requirements and a discussion of the individual parameters in order to avoid comprehension related issues. During the project planning phase, the harmonisation of data should be sufficiently considered. A focus should be placed on a uniform data format and the data acquisition in order to create a sound basis for model comparisons. The use of open and licensed data increases the transparency and reproducibility of model comparisons. In addition, the provision of metadata for model inputs has a favourable cost-benefit ratio. Even without a full harmonisation of input data, can a sound documentation provide a high degree of reproducibility and avoid parameterisation errors.

Acknowledgements

The research for this paper was performed within the projects 'MODEX' supported by the German Federal Ministry of Economic Affairs and Energy under the grant numbers 03ET407 plus the individual project numbers 4-9. The authors thank all project partners for their assistance and participation in the harmonisation process.

Author contribution

Hedda Gardian: Conceptualisation, Methodology, Investigation, Data curation, Writing - Original draft preparation, Supervision, Writing - Review & Editing, Visualization **Jan-Philip Beck:** Conceptualisation, Methodology, Investigation, Data curation, Writing - Original draft preparation, Supervision, Writing - Review & Editing **Matthias Koch:** Conceptualisation, Methodology, Investigation, Data curation, Writing - Original draft preparation, Writing - Review & Editing **Robert Kunze:** Conceptualisation, Methodology, Investigation,

Data curation, Writing - Original draft preparation, Writing - Review \& Editing **Christoph Muschner**: Conceptualisation, Methodology, Investigation, Data curation, Writing - Original draft preparation, Writing - Review \& Editing **Ludwig Hülk**: Conceptualisation, Methodology, Investigation, Data curation **Michael Bucksteeg**: Conceptualisation, Methodology, Investigation, Data curation, Writing - Review \& Editing

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- [1] S. C. Bhattacharyya and G. R. Timilsina, "A review of energy system models," *International Journal of Energy Sector Management*, vol. 4, no. 4, 2010.
- [2] S. Pfenninger et al., "Opening the black box of energy modelling: Strategies and lessons learned," *Energy Strategy Reviews*, vol. 19, pp. 63-71, 2018.
- [3] L. Hülk, B. Müller, M. Glauer, E. Förster and B. Schachler, "Transparency, reproducibility, and quality of energy system analyses – A process to improve scientific work," *Energy Strategy Reviews*, vol. 22, pp. 264-269, 2018.
- [4] M. Wilkinson, M. Dumontier, I. Aalbersberg and et al. , "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, 2016.
- [5] Projektträger Jülich, Forschungszentrum Jülich GmbH, "Modellexperimente für die Energiewende," [Online]. Available: <https://www.energiesystemforschung.de/foerdern/modex>. [Accessed 15 June 2021].
- [6] Stanford University, "Energy Modeling Forum," [Online]. Available: <https://emf.stanford.edu/>. [Accessed 7 June 2021].
- [7] U. Fahl, *Energiemodelle zum Klimaschutz in Deutschland : strukturelle und gesamtwirtschaftliche Auswirkungen aus nationaler Perspektive*, Heidelberg: Physica-Verlag, 1999.
- [8] Technische Universiteit Delft, "Energy Modelling Platform for Europe," EMP-E, 2021. [Online]. Available: <https://www.energymodellingplatform.eu/>. [Accessed 07 June 2021].
- [9] P. Capros, V. Panos, L. Mantzos, M. Zeka-Paschou, V. Papandreou, E. Argiri, M. Uytterlinde, H. Rösler, K. Smekens, G. Martinus, E. v. Thuijl, B. v. d. Zwaan, H. d. Coninck, P. Criqui, S. Mima, P. Menanteau, E. Bellevrat and L. Schrattenholzer,

- "CASCADE MINTS - CAse Study Comparisons And Development of Energy Models for INtegrated Technology Systems," 2004.
- [10] T. C. Johns and et al., "Climate change under aggressive mitigation: the ENSEMBLES multi-model experiment," *Climate Dynamics*, vol. 37, pp. 1975-2003, 2011.
- [11] G. Luderer, E. Kriegler, L. Delsa, A. Arvesen, O. Edelenbosch, J. Emmerling and V. Krey, "Deep carbonisation towards 1,5°C-2°C stabilisation: A synthesis of results from the ADVANCE project," 2016.
- [12] G. Luderer, R. C. Pietzcker, S. Carrarab, H. S. d. Boerd, S. Fujimori, N. Johnson, S. Mima and D. Arent, "Assessment of wind and solar power in global low-carbon energy scenarios: An introduction," *Energy Economics*, vol. 64, pp. 542-551, May 2017.
- [13] H. C. Gils, P. Sterchele, C. Kost, L. Brucker, T. Janßen, C. Krüger, D. Schüwer and H.-J. Luhmann, "RegMex - Modellexperimente und -vergleiche zur Simulation von Wegen zu einer vollständig regenerativen Energieversorgung," 2018.
- [14] H. C. Gils, T. Pregger, F. Flachsbarth, M. Jentsch and C. Dierstein, "Comparison of spatially and temporally resolved energy system models with a focus on Germany's future power supply," *Applied Energy*, vol. 225, p. 113889, 2019.
- [15] Y. Scholz, B. Fuchs, F. Borggreffe, K.-K. Cao, M. Wetzel, K. von Krbek, F. Cebulla, H. C. Gils, F. Fiand, M. Bussieck, T. Koch, D. Rehfeldt, A. Gleixner and D. Khabi, "Speeding up Energy System Models - a Best Practice Guide," 2020.
- [16] G. Savvidis, K. Siala, C. Weissbart, L. Schmidt, F. Borggreffe, S. Kumar, K. Pittel, R. Madlener and K. Hufendiek, "The gap between energy policy challenges and model capabilities," *Energy Policy*, vol. 125, pp. 503-520, 2019.
- [17] E. Kriegler and e. al., "Assessing Pathways toward Ambitious Climate Targets at the Global and European levels: A Synthesis of Results from the AMPERE Project," IIASA Policy Report, 2014.
- [18] OpenEnergyPlatform, "Open Energy Metadata - Open Energy Metadata Description," 2021. [Online]. Available: https://github.com/OpenEnergyPlatform/oemetadata/blob/a37da7c736fb65f3b556608bdc97ad6fd537ac54/metadata/v141/metadata_key_description.md. [Accessed 23 08 2021].
- [19] M. Pritoni, "Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis," *Energies*, vol. 7, no. 14, 2021.
- [20] L. Barbosa and et al., "Practical guide on Open Science for researchers," 2021. [Online]. Available: 10.5281/ZENODO.3968115.
- [21] R. Kunze and S. Schreiber, "Model Coupling Approach for the Analysis of the Future European Energy System," in *The Future European Energy System*, D. Möst et al., Ed., Cham, Springer, 2021, pp. 27-55.

- [22] IEA, World Energy Outlook 2018, Paris, 2018.
- [23] 50Hertz Transmission GmbH, Amprion GmbH, TenneT TSO GmbH, TransnetBW GmbH, Netzentwicklungsplan Strom, Version 2019, zweiter Entwurf der Übertragungsnetzbetreiber, vol. zweiter Entwurf der Übertragungsnetzbetreiber, Berlin, Dortmund, Bayreuth, Stuttgart, 2019.
- [24] ENTSO-E, ENTSOE, TYNDP 2018, Scenario Report, Main Report, Brussels, 2019.
- [25] Boston Consulting Group, Prognos, Klimapfade für Deutschland, München, Hamburg, Basel, Berlin, 2018.
- [26] Umweltbundesamt, Berichterstattung unter der Klimarahmenkonvention der Vereinten Nationen und dem Kyoto-Protokoll 2020, Nationaler Inventarbericht zum Deutschen Treibhausgasinventar 1990 - 2018, Dessau-Roßlau, 2020.
- [27] ENTSO-E, "entsoe Transparency Platform," 2019. [Online]. Available: <https://transparency.entsoe.eu/>. [Accessed 15 June 2021].
- [28] ESA2, "Shaping our energy system – combining European modelling expertise: Case studies of the European energy system in 2050," 2013.
- [29] ESA2, "Database of the REFLEX project in the ESA2 Data Warehouse," [Online]. Available: <https://data.esa2.eu/tree/reflex>. [Accessed 19 08 2021].
- [30] IIASA and the pyam developer team, "pyam: analysis and visualization of integrated-assessment & macro-energy scenarios," IIASA and the pyam developer team, 2021. [Online]. Available: <https://pyam-iamc.readthedocs.io/en/stable/>. [Accessed 31 8 2021].
- [31] VTT Technical Research Centre of Finland Ltd, "Spine," 2021. [Online]. Available: <http://www.spine-model.org/contacts.htm>. [Accessed 31 8 2021].
- [32] Otto von Guericke Universität Magdeburg, RLI, Europa-Universität Flensburg, et al., "https://openenergy-platform.org/about/," 2021. [Online]. Available: <https://openenergy-platform.org>. [Accessed 31 8 2021].
- [33] RLI, et al., "oedatamodel," 2020. [Online]. Available: <https://github.com/OpenEnergyPlatform/oedatamodel>. [Accessed 31 8 2021].
- [34] OpenEnergyPlatform, "Overview - Factsheets," 2021. [Online]. Available: <https://openenergy-platform.org/factsheets/overview/>. [Accessed 23 08 2021].
- [35] German Aerospace Center - Institute of Networked Energy Systems, "Research Project LOD-GEOSS," 2021. [Online]. Available: https://www.dlr.de/ve/en/desktopdefault.aspx/tabid-14162/24548_read-59660/. [Accessed 23 08 2021].
- [36] Reiner Lemoine Institut, "Automated comparison of energy scenarios – SIROP," 2021. [Online]. Available: <https://reiner-lemoine-institut.de/en/automated-comparison-energy-scenarios-sirop/>. [Accessed 23 08 2021].

- [37] D. Huppmann et al., "pyam: Analysis and visualisation of integrated assessment and macro-energy scenarios," *Open Research Europe*, vol. 1, p. 74, June 2021.
- [38] M. Booshehri et al., "Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis," *Energy and AI*, vol. 5, September 2021.
- [39] R. Morrison, "Energy system modeling: Public transparency, scientific reproducibility, and open development," *Energy Strategy Reviews*, vol. 20, pp. 49-63, 2018.
- [40] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016.
- [41] R. Kunze and S. Schreiber, "Model Coupling Approach for the Analysis of the Future European Energy System," in *The Future European Energy System*, D. Möst et al., Ed., Springer, Cham, 2021, pp. 27-51.