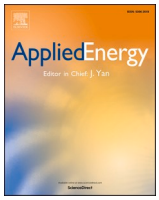
Applied Energy 304 (2021) 117745

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Promoting reproducibility and increased collaboration in electric sector capacity expansion models with community benchmarking and intercomparison efforts

Candise L. Henry [a,b,*], Hadi Eshraghi [c], Oleg Lugovoy [d], Michael B. Waite [e], Joseph F. DeCarolis [c], David J. Farnham [b], Tyler H. Ruggles [b], Rebecca A.M. Peer [f,b], Yuezi Wu [e], Anderson de Queiroz [c], Vladimir Potashnikov [g], Vijay Modi [e], Ken Caldeira [b]

[a] RTI International. Center for Applied Economics and Strategy. 3040 E Cornwallis Rd, Durham, NC 27709, United States
[b] Carnegie Institution for Science. 260 Panama St, Stanford, CA 94305, United States
[c] Department of Civil, Construction, and Environmental Engineering, NC State University. 2501 Stinson Drive Box 7908, Raleigh, NC 27695, United States
[d] Environmental Defense Fund. 1875 Connecticut Ave NW, Ste 600, Washington, DC 20009, United States
[e] Quadracci Sustainable Engineering Lab, Columbia University. 500 West 120th Street, New York, NY 10027, United States
[f] Department of Civil and Natural Resources Engineering, University of Canterbury. Private Bag 4800, Christchurch 8140, New Zealand
[g] Gaidar Institute for Economic Policy. 3-5 Gazetny Lane, Moscow 125993, Russia

HIGHLIGHTS

- Both parametric and structural differences exist in capacity expansion models.
- We used simple cases to eliminate all parametric uncertainty in benchmarking effort.
- We identified structural differences that were previously assumed to be unimportant.
- We introduce an open-source test dataset for the community to use and build on.
- Community-wide benchmarking increases transparency among capacity expansion models.

ARTICLE INFO

Keywords:
Capacity expansion models
Model benchmarking
Energy
Optimization models
Electricity systems

ABSTRACT

Electric sector capacity expansion models are widely used by academic, government, and industry researchers for policy analysis and planning. Many models overlap in their capabilities, spatial and temporal resolutions, and research purposes, but yield diverse results due to both parametric and structural differences. Previous work has attempted to identify some differences among commonly used capacity expansion models but has been unable to disentangle parametric from structural uncertainty. Here, we present a model benchmarking effort using highly simplified scenarios applied to four open-source models of the U.S. electric sector. We eliminate all parametric uncertainty through using a common dataset and leave only structural differences. We demonstrate how a systematic model comparison process allows us to pinpoint specific and important structural differences among our models, including specification of technologies as baseload or load following generation, battery state-of-charge at the beginning and end of a modeled period, application of battery roundtrip efficiency, treatment of discount rates, formulation of model end effects, and digit precision of input parameters. Our results show that such a process can be effective for improving consistency across models and building model confidence, substantiating specific modeling choices, reporting uncertainties, and identifying areas for further research and development. We also introduce an open-source test dataset that the modeling community can use for unit testing and build on for benchmarking exercises of more complex models. A community benchmarking effort can increase collaboration among energy modelers and provide transparency regarding the energy transition and energy challenges, for other stakeholders such as policymakers.

* Corresponding author at: RTI International. Center for Applied Economics and Strategy. 3040 E Cornwallis Rd, Durham, NC 27709, United States.
E-mail address: clhenry@rti.org (C.L. Henry).

https://doi.org/10.1016/j.apenergy.2021.117745
Received 16 May 2021; Received in revised form 23 August 2021; Accepted 27 August 2021
Available online 13 September 2021

1. Introduction

Models focused on the electric sector often use different cost and technical inputs, demand and resource availability profiles, temporal and spatial resolutions, capacity retirement and policy assumptions, and even mathematical formulations in order to address research questions of various scales and purposes. It is unsurprising, therefore, that different models output different results for a given planning or operations problem. For instance, to identify where new transmission would have the greatest impact on overall grid stability, sub-hourly resolution models meant for high-fidelity power flow scheduling will likely not produce the same results as models aimed at understanding the century-scale transition to a low-carbon economy. But even among models that are designed to answer the same type and scale of questions, dissimilar results are common given the parametric and structural differences among competing models [1,2]. Parametric uncertainty arises when researchers use diverging input parameters such as costs, technology specifications (e.g., heat rates), electricity demand, and resource availability. Structural uncertainty, on the other hand, stems from differences in model representations of electric systems and how models are mathematically formulated. Specific examples in which model structures can differ include whether long duration storage is included in a capacity expansion model or whether battery roundtrip efficiency is applied to electricity entering or exiting the battery. Given that both parametric and structural differences are often present across models, it can be difficult to identify the origin(s) of diverging results among even seemingly similar models. Yet, in order to fully understand why models intended to address similar questions might give different outcomes, it is important for researchers to disentangle parametric from structural uncertainty.

Here, we present the first intercomparison effort of capacity expansion models that removes all parametric uncertainty to focus only on structural differences. Our approach allowed us to pinpoint the origins of specific divergences between our models – also applicable to other capacity expansion models – that otherwise would have been overlooked or perceived as unimportant. More importantly, the structural differences we identified here indicate the potential for significant divergences among the more complex model formulations common in the research literature and planning studies. As such, this work highlights the benefits of a concerted model benchmarking effort across the broader capacity expansion modeling community. Systematic approaches for comparing models could help energy researchers identify how the assumptions about inputs and model structures embedded in their own models differ from others that are designed to answer similar research questions. They could help researchers justify the assumptions they make in their own models in the context of other models. Furthermore, such approaches enable new forms of communication among modeling efforts that could help further strengthen the growing community of energy modelers, and lead to the continued improvement of individual models. A systematic model benchmarking effort can also benefit the broader energy research community by increasing the transparency of models being used in power sector planning and policymaking. It is helpful for stakeholders to understand the assumptions, strengths, limitations, and uncertainties of these models, as they are often used to help inform the direction of power sector development.

While we focus on four open-source models that represent the U.S. electric sector, we note that our results should be relevant to other optimization-based capacity expansion models that represent other countries. Capacity expansion modelers across the world can benefit from community benchmarking efforts.

1.1. Electric sector capacity expansion models

In this work, we specifically focus on electric sector capacity expansion models, which are used in academic studies and in policymaking today to evaluate power sector policies and project the economic viability of various technologies [3–13]. We define capacity expansion models as those that employ linear or mixed integer programming to examine capacity deployment and utilization over future decades. These models are often initiated from baselines that represent real existing infrastructure, but can begin from baselines with no existing infrastructure. Moreover, we define capacity expansion models as those that optimize at an aggregated level without chronological unit commitment, which excludes those that are coupled with production cost models to resolve the behavior of individual power plants [14]. The capacity expansion models we examine here consider electricity transmission but ignore details about power flow parameters such as voltage angles, real power, and reactive power.

Our model simulations assume perfect markets and rational actors with complete information and perfect foresight. The relative simplicity of these models should increase likelihood of different models finding problem solutions that are numerically similar. Of course, electricity systems around the world operate under different and evolving degrees of regulatory control and market forces. In contrast to perfect foresight, uncertainty in prices and demand are expected to increase due to increased competition that accompanies the trend towards deregulation [15]. Linear optimization cannot capture the richness of decisions as they are made, so utility planners are moving from optimization models to behavioral simulation modeling and agent-based methods [16]. Here we focus on the degree of agreement among the simplest and most transparent electricity system models, with the understanding that this work would help provide basis for similar comparison of more comprehensive models.

1.2. Existing energy system model comparison efforts

Various projects to compare energy system models exist. The Stanford Energy Modeling Forum represents one of the model intercomparison efforts aimed at answering specific policy questions [17]. Their primary goal is not to benchmark and identify differences among models, but to produce an ensemble of projections from different integrated assessment models (IAMs) and energy models given prescribed policy scenarios. Several other past and current US and EU sponsored projects have taken a similar approach for the intercomparison of IAMs, including the Innovation Modeling Comparison Project [18], Climate Change Science Program, Asian Modeling Exercise [19], RoSE Project [20], and Latin America Modeling Project [21].

Similarly, Long et al. [22] explored the costs and capacity requirements of various scenarios under California's new zero carbon mandates using three distinct models with the same historical input data and future technology costs. The goal of this work was to create and assess an ensemble of projections rather than to benchmark the models used.

At the same time, many authors have conducted comprehensive reviews on the landscape of electricity sector models and the modeling tools available to address different questions [23–28]. However, these works are generally aimed at qualitatively comparing the capabilities of various existing capacity expansion models rather than conducting model intercomparison or benchmarking efforts with harmonized input datasets.

The Renewable Energy and Efficiency Modeling Analysis Partnership (REMAP) in 2009 was an early intercomparison effort that attempted to benchmark what the authors classified as energy-economy models using harmonized assumptions [1]. The developers of eight models, which included capacity expansion models, collaborated to assess whether model outputs would converge if major input assumptions were consistent across the models. However, due to the complexity of some of the models and the fact that they were burdensome to change, inputs could not be completely harmonized and both parametric and structural uncertainties remained in the outcomes. The REMAP authors did find, though, that variation among models diminished significantly when the inputs were better aligned.

Mai et al. [2] conducted a similar, more recent benchmarking effort for electric sector capacity expansion models. The authors ran a set of coordinated scenarios with harmonized technology costs on three widely used models, which each have highly resolved representations of the U.S. electricity sector. Their primary aims were to identify best practices for representing variable renewable (VRE) technologies and to isolate the effect that model structures might have on VRE deployment outcomes. They found that significant differences in wind and solar installed capacities in the least-cost system persisted in their comparison, implying that the parametric and structural components they did not harmonize – such as transmission costs, financing costs, and model structures – had important effects on model outputs.

Aside from these individual studies, community-wide benchmarking efforts for capacity expansion models and broader macro-energy system models [29] do not exist at the same scale as they do for production cost (unit commitment) models. The Institute of Electrical and Electronics Engineers (IEEE) developed Benchmarks for the Unit Commitment Problem in Power Grid Lib (PGLib), which is a collection of open-source GitHub repositories containing benchmarks for different production cost models to validate power system algorithms [30], (power-grid-lib. github.io). Research groups are encouraged to input the various case files into their own models to compare results against the benchmark values. This approach not only allows model developers to better understand how their model differs from other models, but it also helps developers maintain consistency when expanding or improving existing models.

In this study, we highlight the benefits of a benchmarking process for electric sector capacity expansion models by demonstrating that specific and important structural differences can be identified through this process. We focused on four models that have all been used in published literature. This work differs from previous studies because we conducted a harmonization process of "turning off" various model capabilities in order to remove all parametric uncertainty from these models. This included eliminating transmission capabilities, capacity retirements, operating and capacity reserves, frequency regulations, policy constraints, and any technologies other than solar, wind, natural gas, nuclear, and battery storage. Capital, operation and maintenance, and fuel costs as well as technical specifications of these five technologies were aligned using a common dataset. This allowed us to pinpoint specific structural differences in how these five technologies and the objective function are formulated, without the compounding uncertainty from parametric differences [31]. Identifying where structural differences can occur between models is important because it allows model users to determine which structural assumptions should be considered alongside parametric input decisions when addressing certain research questions. The purpose of this work is to demonstrate the advantages of model benchmarking for electric sector capacity expansion models, to identify important structural differences across the participating models, and to provide a publicly available input dataset that can serve as the basis for the development of a suite of official tests. While the work presented here focuses on the electric sector, the approach can be applied more broadly to a variety of macro-energy models.

2. Methods

2.1. Model structures

In this analysis, we compared four models with similar core cost optimization structures and system constraints: North Carolina State University's Tools for Energy Model Optimization and Analysis (Temoa), Carnegie Institution for Science's Macro Energy Model (MEM), the independently-developed energyRt model used by the Environmental Defense Fund, and Columbia University Sustainable Engineering Lab's System Electrification and Capacity Transition (SECTR) model. All co-authors of this paper were involved in the development of one of the four models. The respective authors for each model, the high-level model details, and links to each model's Github repository can be found in Table 1.

These four models are open-source and are all designed to minimize total system cost given cost and technical specifications, capacity factors representing resource availability, and a time-series dataset of electricity demand. The objective function underpinning each of the models can be described by:

$$\textbf{\textit{minimize}}: \text{system cost} = \sum_g p^g_{\text{fixed}} C^g + \sum_g \left(\frac{\sum_t p^g_{\text{var}} D^g_t}{T} \right) + \sum_s p^s_{\text{fixed}} C^s + \sum_s \left(\frac{\sum_t p^{\text{ch}}_{\text{var}} D^{\text{ds}}_t}{T} \right) + \sum_s \left(\frac{\sum_t p^{\text{ds}}_{\text{var}} D^{\text{ch}}_t}{T} \right) \quad (1)$$

where:

g is the generation technology

s is energy storage

ch is electricity charging (entering) the battery

ds is electricity discharging (exiting) from the battery

p_{fixed} is the capital and fixed O&M cost[1] ($/kW for generators, $/kWh for storage)

p_{var} is the variable cost including fuel and variable O&M ($/kWh)

C is the capacity of the technology (kW for generators, kWh for storage)

D_t is the electricity dispatch at time step t (kW)

T is the total hours in the simulation

Here, the decision variables for determining minimum system cost are the capacities (C^g, C^s) and hourly dispatch (D^g_t, D^{ch}_t, D^{ds}_t) of each technology. The constraints for capacity and dispatch of each generation and storage technology as well as for system energy balance are:

$$C^{g,s} \geq 0 \quad \forall g, s \quad (2)$$

$$0 \leq D^g_t \leq C^g f^g_t \quad \forall g, t \quad (3)$$

$$0 \leq D^{\text{ch}}_t \leq \frac{C^s}{\tau^{ch}} \quad \forall s, t \quad (4)$$

$$0 \leq D^{\text{ds}}_t \leq \frac{C^s}{\tau^{ds}} \quad \forall s, t \quad (5)$$

$$0 \leq S^s_t \leq C^s \quad \forall s, t \quad (6)$$

$$0 \leq D^{\text{ds}}_t \Delta t \leq S^s_t (1 - \delta) \quad \forall s, t \quad (7)$$

$$\sum_g D^g_t \Delta t + \sum_s D^{\text{ds}}_t \Delta t = M_t + \sum_s D^{\text{ch}}_t \Delta t \quad \forall g, s, t \quad (8)$$

[1] Note that the fixed O&M costs are amortized in these models.

Table 1
Description of the four models used in this study.

Model	Programming Language	Optimizer	Default Time Steps	Geographical Coverage	Commodities	Organization	Authors	Repository
Temoa	Python	Any optimizer	From sub-hourly to seasonal	Regional to National	Whole energy system energy carriers and emissions	North Carolina State University	HE, JD	github.com/TemoaProject
MEM	Python	Any optimizer Default: Gurobi	Hourly	Regional to National (idealized)	Electricity	Carnegie Institution for Science	CLH, DJF, THR, RP, KC	github.com/carnegie/MEM_public/tree/Henry_et_al_2021
energyRt	R + (GAMS/ Python/ Julia/ MathProg)	Any optimizer	From sub-hourly to seasonal	Regional to National	Any	Independent development	OL, VP	github.com/energyRt/energyRt
SECTR	Python	Gurobi	Hourly	Regional to National	Electricity Fossil fuels in buildings and vehicles	Quadracci Sustainable Engineering Lab, Columbia University	MBW, YW, VM	github.com/SEL-Columbia/SECTR-OneNode

where:

f_t is the capacity factor of the generation technology in time step t (fraction)

τ is the storage charging and discharging duration (h)

S_t is the energy in storage at the end of time step t (kWh)

δ is the storage decay rate (fraction)

M_t is electricity demand at time step t (kWh)

Δt is time step size (h)

While all four models have this underlying core formulation capability, each also incorporates additional components that increase the complexity of the model and could potentially create differences in model structure. For instance, the four models simulate different mixes of technologies (e.g., different types of storage), and have different rules for separate technologies and how they behave (e.g., ramping limits or reserve margin constraints). In this analysis, we attempted to reduce the presence of components that can lead to structural complexity by "turning off" model features where possible. Specifically, each model was simplified to operate as a one-node electricity system where 100% of demand is met by an electricity supply at the same node without any transmission costs or losses. We simulated only one year of hourly dispatch from a greenfield site (no existing infrastructure) to remove structural differences from capacity building and vintaging constraints. Subsequently, decommissioning costs are also not considered. We did not specify upper bounds on the quantity of new generation that can be built. We did not include operating reserve, capacity reserve margin, frequency regulation, or policy constraints. Finally, we specified the technology mix for each case we ran as well as the cost inputs and technical specifications for these technologies. Such parsimonious assumptions help to avoid strong priors about what model developers expect from capacity and generation mixes in the real world.

2.2. Harmonized data inputs

We created five hypothetical cases with harmonized input datasets for which the models solved for the cost-minimizing system (Table 2):

S1: Natural gas, nuclear, wind, solar, and storage at costs defined in AEO 2018 (Base Case)
S2: Natural gas, nuclear, wind, solar, and storage at costs that were predefined fractions of AEO 2018 costs such that all technologies are cost competitive (more details below) (Alternative Case)
S3: Solar and battery only at AEO 2018 costs
S4: Wind only at AEO 2018 costs
S5: Nuclear only at AEO 2018 costs

In each of these cases, all models used the same capital costs, fuel costs, and fixed and variable operation and maintenance (O&M) costs from the Energy Information Administration (EIA) 2018 Annual Energy Outlook (AEO) for each of the generation and storage technologies, as well as the same discount rate, capital recovery factors, and project lifetimes (Appendix Table A1). Costs were held constant through each model run, i.e., costs of generation did not change with increasing capacity. Battery charging efficiency, decay rate, and charging time for storage were specified. The same hourly demand profile from 2016 was used by all models and represents aggregated demand for the contiguous U.S.[32]. The same 2016 hourly solar and wind capacity factor profiles were also applied to all models, and were derived from wind speed and solar irradiance information from the MERRA-2 satellite reanalysis dataset [33]. The models solved all 8784 h simultaneously as the solution approach. The harmonized cases and input datasets described here allow us to benchmark the outputs of our different models but are not meant to represent predictions of future conditions. As such the years represented by our datasets are unimportant for the purpose of this analysis. The input datasets for the five cases used can be accessed at github.com/carnegie/capacity-expansion-model-intercomparison.

Case S1 was selected for model benchmarking because it applies AEO costs for multiple technologies that were meant to reflect real prices in 2018. This case provided a starting point for us to begin comparing model structures before assessing individual technologies. In case S1, the capital cost of natural gas combined cycle plants is almost half that of

Table 2
Single-node cases used in this model intercomparison. The capital, fuel, and O&M cost inputs are taken from the [34] Annual Energy Outlook. The hourly demand profile represents aggregate demand for the contiguous U.S. [32] and the hourly solar and wind capacity factor profiles are derived as a U.S.-wide average from MERRA-2 [33]. The input datasets are available at: github.com/carnegie/capacity-expansion-model-intercomparison.

Case	S1	S2	S3	S4	S5
Technologies	Natural gas, nuclear, solar, wind, and battery storage		Solar and battery storage	Wind	Nuclear
Cost Inputs	2018 EIA AEO	Hypothetical costs adjusted from 2018 EIA AEO	2018 EIA AEO for solar and battery	2018 EIA AEO for wind	2018 EIA AEO for nuclear
Demand Input	Hourly demand for one year [32]				
Solar Input	Hourly solar capacity factor for one year [33]				
Wind Input	Hourly wind capacity factor for one year [33]				

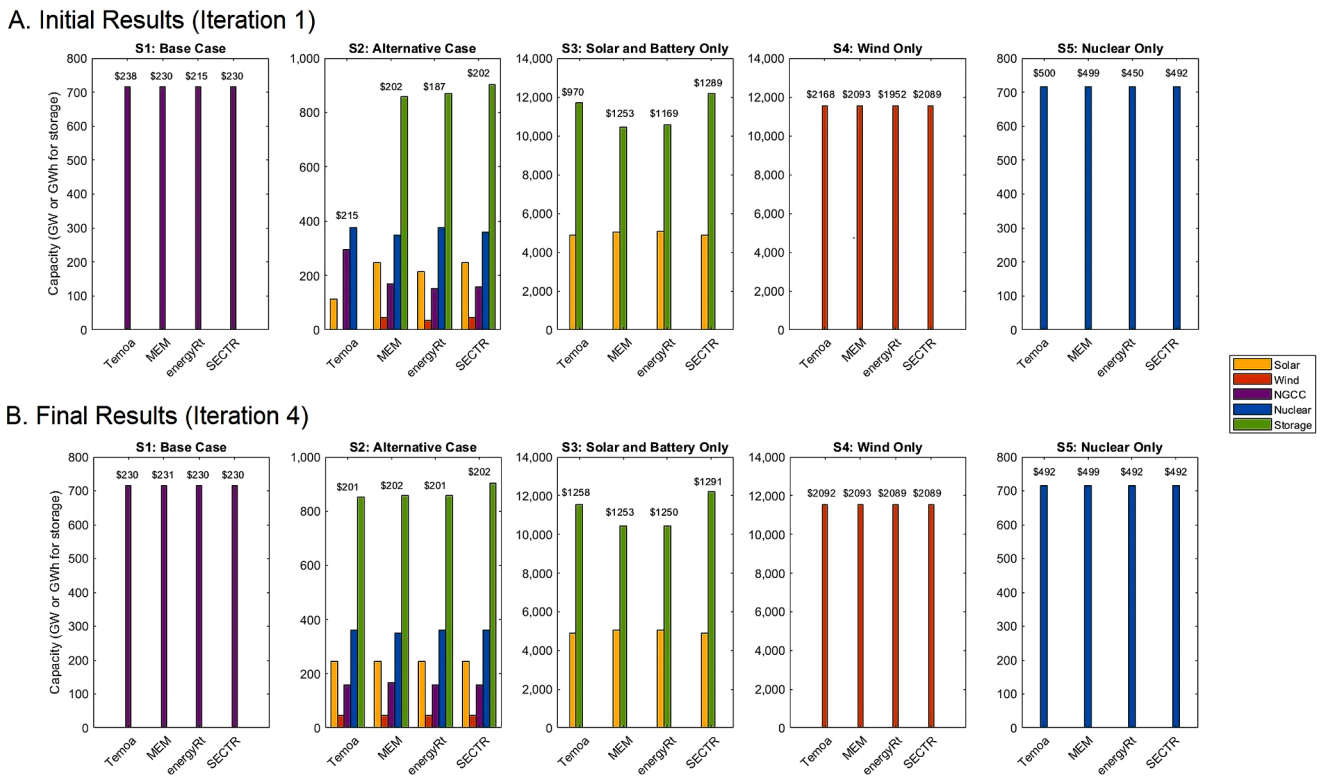


Fig. 1. **(a)** Initial and **(b)** final built generation capacities for each of the four models in each of the five cases. The units are in GW for natural gas, solar, wind, and nuclear capacities and GWh for battery storage. As can be seen in the y-axes, the built capacities for all models in case S3 and S4 are a factor of 10 greater than those for cases S1, S2, and S5. Also shown above each model are the corresponding total annual system costs (in billion dollars US) for each of the cases. Note that this is a 1-year simulation, so the total annual system cost equals the total system cost.

the next cheapest technology, making it the most cost competitive of the five technologies we include. In all five cases, costs of curtailing wind and solar were zero.

The natural gas, nuclear, wind, solar, and storage cost inputs in case S2 were selected to capture a scenario where all technologies are cost competitive. These costs are not reflective of any real-world scenario but were determined by the authors using one of the models, MEM. Specifically, we iteratively modified the AEO 2018 case S1 costs until all technologies were deployed at some point during the 8784-hour time frame. Each modeling group then exogenously applied these same exact predetermined technology costs into their respective models. This case was designed to increase complexity in our highly simplified system and highlight possible differences in each technology between models.

Cases S3-S5 were designed for unit testing of mathematical formulations representing individual technologies. Unit tests are a form of software testing where individual components of a code are tested to validate that each unit performs as expected. Here, our unit tests isolated individual technologies in our models and allowed us to pinpoint where structural differences exist in our code.

While the cases introduced here were developed over time through dialogue among the four modeling groups involved in this paper, they can be applied to other models to investigate and benchmark results. Moreover, these cases were designed to be simple unit tests that can be expanded to test other components of capacity expansion and macro-energy models.

2.3. Comparison metrics

In order to compare the models, we examined their output system costs, generation and storage capacities, and hourly dispatch profiles for each of the cases. The combination of system cost and generation and storage capacity allowed us to determine whether model structures differed enough to warrant further investigation. Then, to pinpoint the causes of any differences in results and to identify which technologies led to these differences, we investigated the hourly dispatch profiles for more detail. The hourly dispatch profiles allowed us to hone in on differences in the exact behavior of the generation technologies and the hourly charging and discharging of the energy storage asset.

2.4. Iterative process of model structure harmonization

After the initial model runs with the harmonized parameters were completed, we aimed for further convergence of results through harmonization of model structure. Our aim was not to prescribe certain model structures as being more or less representative of real systems, but rather to pinpoint which structural differences impacted results and to initiate discussion within our group about why certain implementation choices were made. Since our four models were already simplified through the "turning off" of various components, the additional level of effort for further harmonization was not prohibitive. We recognize that it may be unrealistic for more complex models to harmonize to the degree that we do here for intercomparison, but we also emphasize how useful this benchmarking process is for model development by demonstrating the type of information that we can learn from systematic harmonization.

Though this model structure harmonization process is part of the methodology, it is presented in the Results and Discussion section below because it was an iterative and cooperative effort that took several steps with intermediate results. Between each iteration, the four model groups met to discuss model outputs and to collectively identify the potential sources of differences, as well as to come to agreement on synthesizing model structures.

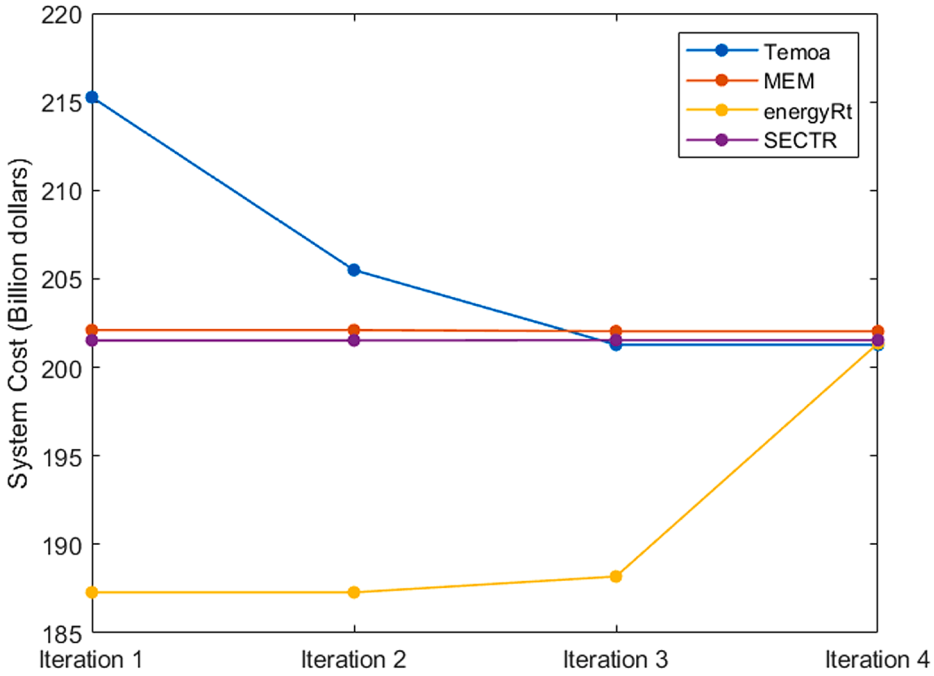


Fig. 2. Each model's case S2 annual system cost after every iteration of model calibration, showing how the models converge with each adjustment. We use case S2 as the example here because this case yielded the greatest initial (iteration 1) differences in both system cost and built capacity.

3. Results and discussion

3.1. Initial and final model results

Fig. 1 shows the initial and final model results for all four models. The initial results were produced using the harmonized input datasets from cases S1-S5 with various model components "turned off", but without iterative model structure harmonization (described in the following section). As can be seen in Fig. 1A, the built capacities across models in our "starting point" were already similar in cases S1, S4, and S5, due to this initial parameter harmonization process. For instance, for case S1, all models produced the same 716.7 GW generation capacity from natural gas only (Fig. 1), where 716.7 GW corresponds to the maximum hourly demand in our demand time series. Likewise, for cases S4 and S5, all models produced identical generation capacities of wind and nuclear, respectively (Fig. 1).

Despite the similarities in the initial outputs, the system costs across all five cases as well as the built capacities in cases S2 and S3 indicated the presence of structural differences between the models (Fig. 1A). Through our iterative structure harmonization process, we were able to better align model outputs to reach the final values shown in Fig. 1B. For case S1, system costs differed by 10% between models in the initial output and were harmonized to be within 0.3% in the final output. For cases S4 and S5, system costs differed by 11% and 12% between models in the initial output and were harmonized to be within 0.2% and 0.3% in the final output, respectively. Meanwhile, in case S2, both the generation capacity and total system cost evolved substantially between the initial and final iterations (compare Fig. 1A and 1B). The system costs of the four models differed by 13% in the initial output and eventually converged with each structural adjustment to be within 0.3% of one another.

The final output from case S3 (Fig. 1B) reveals that differences remained in the models that could not be addressed through the structural changes we made here. While the final system cost in case S3 differs by less than 3% between models (Fig. 1), the generation capacities in case S3 still differ substantially (16% for battery and 3% for solar) despite parametric and structural harmonization. Potential reasons are discussed in more detail in Section 3.3.

The exact values shown in Fig. 1 can be found in the Appendix (Tables B1 and B2).

3.2. Model structure harmonization process

In order to reach the final aligned model results shown in Fig. 1B, we took several steps to calibrate each of our respective models. These alterations were intended to eliminate model differences identified during the comparison process and thereby help us better understand and identify previously unknown model differences. Since initial differences were most evident in case S2, we use it here to demonstrate our process for identifying structural differences between our models (Fig. 2). The calibration process described below occurred in four steps, with each iteration proceeding after we identified a key source of difference between models. It is important to note again that we are not advocating for the specific calibration choices we made here; in fact, in many of the instances discussed below, the original formulations were modeling choices made by the authors to answer specific questions about the real-world electricity sector. Rather, we describe the harmonization process in order to demonstrate how different choices can lead to differing results.

The only structural alterations in the first (initial) iteration involved "turning off" model components for which we had no harmonized input datasets. These were described in the Methods section. The results can be seen in Fig. 1A and iteration 1 of Fig. 2.

To better align case S2 outputs for the second iteration, the treatment of the dispatchability of nuclear power was calibrated in one of the models: Temoa initially set nuclear as baseload generation that cannot be ramped up and down across times of day, but can shift output from one season to another. Meanwhile, the other models treated nuclear as dispatchable generation similar to natural gas. This difference in load following capability caused nuclear generation to behave differently between Temoa and the other models, thus impacting the generation portfolio and system cost in case S2. This difference could also have been detected by examining the case S5 hourly generation profile of each model after iteration 1. Examining only the system cost and generation capacity of nuclear from case S5 may not have alerted us to this structural difference because the capacity would still be the same,

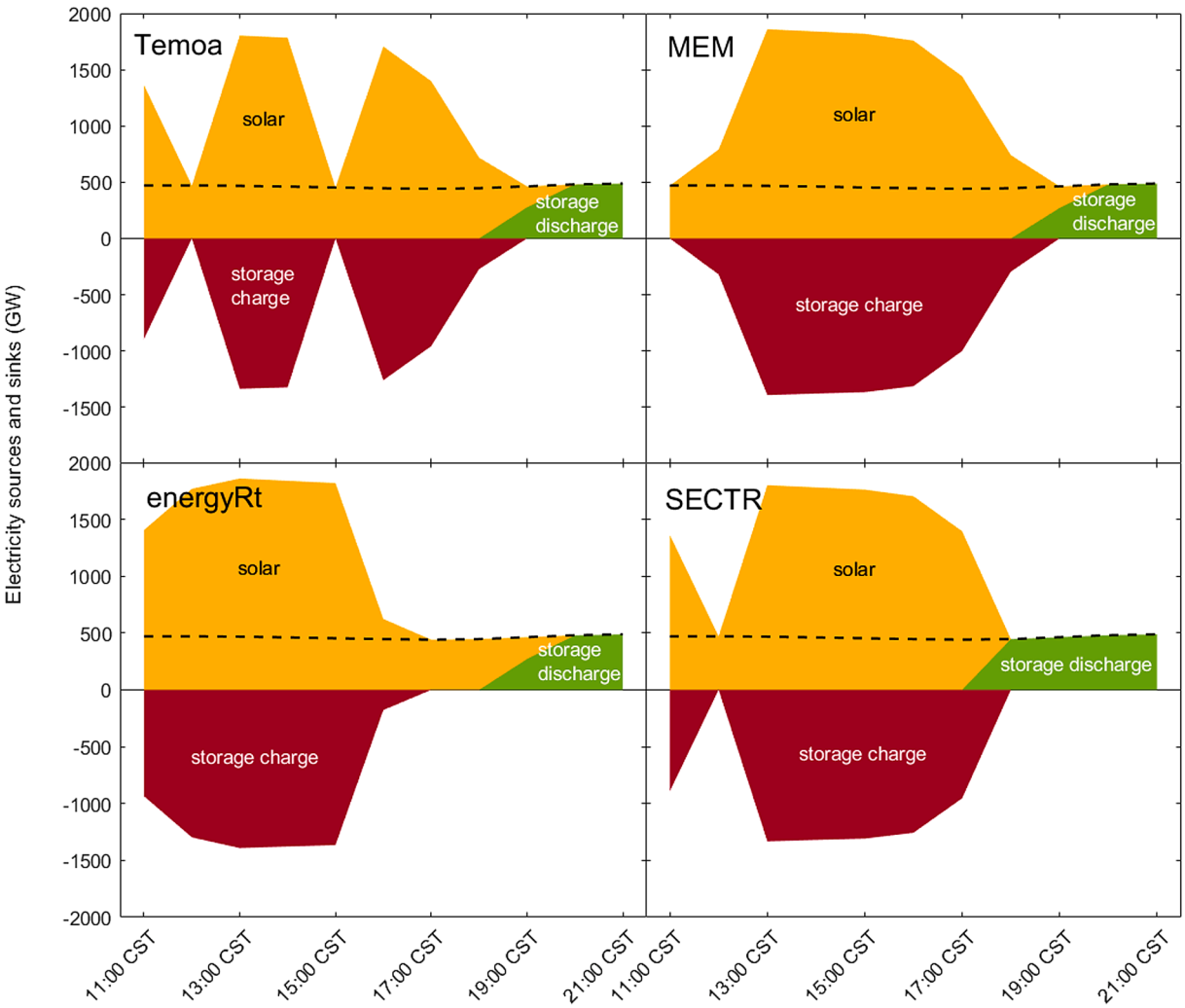


Fig. 3. The same representative ten-hour time series from each model showing the hourly dispatch by each technology under case S3. The positive values (solar and storage discharging) show electricity generated to meet demand, while the negative values (storage charging) show excess electricity generation used to charge the battery. The black dashed line marks demand and appears flat due to the wide y-axis range but is in fact not constant. As can be seen, each model meets 100% of demand at all hours (a requirement specified in the case) but displays a different charging and discharging pattern due to the zero variable costs of solar and storage, which creates non-uniqueness in the optimization decision space. Note that curtailment of solar occurs across all of the models in these hours but is not explicitly shown in this figure.

corresponding to peak demand. This is why examining multiple cases is helpful for a model intercomparison effort. The final results in Fig. 1 show the results after ramping constraints on nuclear were removed from all models, and nuclear was allowed to behave as dispatchable generation across all four models.

After changing the dispatchability of nuclear, we still observed differences in model-selected generation capacity among models in case S2. We pinpointed battery constraints as the source of structural uncertainty because case S3 also yielded differences. While two of the models initially required the battery state-of-charge to be the same at the start and end of the simulation, one model set the battery charge at full capacity in the first hour and empty in the last hour. The fourth model constrained the initial and final battery state-of-charge to be 50%. In the third iteration, models were adjusted so that all had the constraint that the initial and final state-of-charge are equal, model-determined values.[2]

Along the same lines, another source of difference adjusted in the third iteration was related to how battery storage roundtrip efficiency was applied. Roundtrip efficiency can be applied to energy entering the battery (charging), to energy leaving the battery (discharging), or split between charging and discharging. This application impacts the model-selected battery capacity, which in turn impacts the size of the various generation technologies and the system cost. We harmonized all four models in this iteration such that roundtrip efficiency was applied to energy entering the battery.

After the third iteration, all four models yielded similar built capacities in case S2 (Fig. 1B). However, energyRt still had ~7% lower system cost than the others in this case (see iteration 4 in Fig. 2). We found that a substantial difference in system cost formulation involves the treatment of the time value of money and discount rates. When determining present value, it is possible to apply a discount rate at any point in a model year. Meanwhile, some models ignore the time value of money when computing dispatch. By re-parameterizing energyRt to apply the discount rate at the end of the simulation year, system costs converged with the other models in all cases (Figs. 1 and 2). This yielded the final model outputs, iteration 4, shown in Fig. 1B.

[2] Specification of initial and final battery storage amounts can have a greater influence on results for cases with small amounts of simulation time or for models that use representative time slices.

3.3. Remaining model differences

Even with the adjustments discussed above, differences persisted in the computed system costs, built capacities, and hourly dispatch profiles of the four models in all five cases, indicating that underlying model differences remained. There are numerous possible explanations for such deviations; here we described some based on our intercomparison process and modeling experience.

First, differences can persist if there is not a unique solution in the decision space to the cost minimization problem. This can occur, for instance, when the variable costs of generation for wind, solar, and battery are zero as they were in this analysis. The models are indifferent to variations in the precise timing of storage dispatch as long as they do not affect the dispatch cost (and thus the objective function), so there appears to be some randomness to when charge and discharge occurs. Fig. 3 shows an example of this by presenting the hourly dispatch profile from case S3 for a randomly selected 10-hour segment in our simulation year. In all four models, solar generation (yellow block) was used to meet 100% of demand (black dashed line) at the onset of the 10 h. Excess solar in these hours was used to charge the battery (red block), which in turn was dispatched in the later hours (green block) to meet demand when solar generation was not available (i.e., evening hours). However, the models differed as to when and for how many hours charging occurred, as long as battery capacity was sufficient to meet demand requirements throughout the night. If we were to take snapshots of other 10-hour periods, we would see distinct battery charging and discharge behavior among other models. This is important to note because sometimes researchers attempt to draw operational lessons from capacity expansion models, but this example indicates that this use-case must be approached with great care in this class of models (particularly if one is interested in storage operation).

Second, the precision with which model parameters are specified can create differences in model outputs. This class of models often shows "knife-edge" results, where an incremental shift in costs makes one technology cheaper than another, resulting in a step-function change in capacities in the least cost system. We saw a dramatic manifestation of this in case S2, where costs were intentionally selected to be near these edges. We ran case S2 in MEM with three levels of rounding precision: using cost input parameters with float precision, rounded to four decimal places, and rounded to two decimal places. We found that total system costs were $201.9 billion, $202.4 billion, and $210 billion, respectively. This is a 4% increase in total system cost (in dollars) when we move from using costs (in $/kW and $/kWh) with float precision to two decimal places of rounding. Meanwhile, in this knife-edge case, 350 GW of nuclear was built when inputs had float precision but no nuclear was built when inputs were rounded to two decimal places. Such wide variation has potentially major implications for the reliability of model-based planning without sufficient sensitivity analysis.

Relatedly, as with all linear programs, a wide range of matrix coefficient values or very small values can create numerical issues and, potentially, an infeasible model [35]. In this study, no such issues occurred, and all research groups used the same model structure. Consistent parameter scaling across models can help future model intercomparison efforts avoid such issues.

One difference that we identified among our models that did not have a large effect on our results was the formulation of the objective function. Even though all of our models are consistent with Eq. (1), each applied a different method for considering the "end effects" that occur because new capacity costs incurred towards the end of a model time horizon can skew results. One way to formulate this is to explicitly include salvage costs in the objective function, while another is to truncate annual capital payments extending beyond the model time horizon. These different objective function formulations involve distinct assumptions that can cause diverging results but can be difficult to calibrate between models without reformulating the models themselves. Therefore, we did not harmonize the models to address these differences. In Fig. 1, divergences in the objective function formulation are most evident in the wind only (S4) and nuclear only (S5) cases, where the four models build the same amount of wind or nuclear capacity, respectively, but produce different total system costs. However, even without calibrating our models to account for this, the differences in system costs were less than 1% in these two particular cases, so the impacts of the end effects formulation do not appear to have substantially impacted our results. This might be a more important consideration for models that optimize with representative time slices, rather than hourly dispatch, because boundary conditions play a more important role in determining the optimal solution.

3.4. Qualitative assessment of the value of benchmarks and unit testing standards in capacity expansion modeling

Previous work has demonstrated that technology costs and model structures can both play important roles in determining capacity expansion model outcomes [1,2]. However, the parametric and structural uncertainties in models are difficult to disentangle because it is often impractical to fully harmonize both model inputs and structures in intercomparison efforts. Based on the conclusions of previous work, a hypothesis going into this study was that remaining outcome differences in existing intercomparison efforts (such as those by [1] and [2]) emerged primarily from the complex – and thereby difficult to harmonize – structures of the models being studied. We assumed that parsimonious models, where most model components are "turned off", with aligned input costs and technical specifications would yield nearly identical results. We found that, while technology costs no doubt play a central role when optimizing a least cost system, technology-specific structural assumptions and details regarding how models are initialized can substantively influence results. More importantly, we found that having a systematic process for model testing allowed us to pinpoint differences between models that we either did not previously recognize or did not expect to have substantial impact on results. Differences in model outputs thus cannot be blamed on a nebulous "complexity." As such, we emphasize the value of coordinated benchmarking and unit testing efforts among the capacity expansion modeling community for building confidence in model results, substantiating specific modeling choices, reporting uncertainties, and identifying areas for further research and development.

While it is impractical to expect last-digit agreement among energy system optimizations conducted using different models, similarity in quantitative outputs and qualitative agreement among models is essential to their reliability as planning tools. Confidence in model results can be increased through analysis of discrepancies between models. Given the proliferation of such models and the diversity of their developers and applications, a systematic approach for capacity expansion models would be necessary to build the same confidence achieved across a wide diversity of climate models [36] and building energy models [37]. A systematic approach includes common input datasets for verification and benchmarking as well as standard processes to step through comparison efforts.

Relatedly, the results of energy system models are typically reported without any associated uncertainties. For results to be more useful, the energy modeling community should identify and be transparent about the parametric and structural contributors to uncertainties in model results. In cases where uncertainties are hard to quantify, sensitivity analyses that illustrate the dependencies of model results on reasonable variation of input parameters could be useful for estimating parametric uncertainties.

It is important to point out that differences in model outputs – if adequately documented and explained – can be an asset to the broader energy community because they reflect real differences of opinion between model users about cost projections, technological assumptions, and more. Diverging model results should not invalidate the models. However, a coordinated benchmarking effort could help researchers

identify a model's coding errors and its non-erroneous fundamental differences from other models, as well as areas for further research and development.

The cases and input datasets we introduce here can serve as the foundation for further development of a library of official tests to be used for benchmarking capacity expansion and macro-energy models. We will archive these test cases and our benchmark results in a public repository for broader accessibility and to provide opportunities for further development by the energy modeling community. These five cases are good starting points for further development because they apply limited assumptions to a simple system topology, and allow for (1) small computation times (on the order of seconds for our four simple models), (2) easy identification of input-related and non-input-related model discrepancies, and (3) the potential to be expanded both in terms of the number of simple cases and the complexity of existing cases for testing more assumptions. For instance, our cases are built on greenfield sites and simulated over only one year, so do not explore differences in retirement capacity. This shortcoming can be readily addressed with a case that uses a longer time period and a case that includes existing capacity in order to test the treatment of new and existing capacity. Other extensions to the unit tests presented here could include additional electricity generation technologies, multiple nodes with transmission infrastructure, reserve planning, policy-based targets, technologies in other sectors, and other meteorological and hydrological year profiles.

4. Conclusion

Here, we presented the first intercomparison effort of capacity expansion models that removes all parametric uncertainty to focus only on structural differences. We showed how a systematic process for eliminating parametric uncertainty and reducing structural discrepancies can identify structural differences in four electric sector capacity expansion models with the same mathematical definition. By systematically comparing our four models using a set of five test cases and an iterative process involving communication among members from all four model groups, we were able to identify differences that we had not previously recognized would substantially impact our results. The specific structural differences we pinpointed through just four iterations of model harmonization that are likely applicable to other capacity expansion models include:

1. The specification of various technologies as baseload or load following generation.
2. Constraints on the state-of-charge for storage at the beginning and end of the modeled period.
3. The application of battery roundtrip efficiency on energy entering vs. exiting the battery.
4. The treatment of discount rates and time value of money.
5. The non-unique decision space in a cost optimization problem that can be created with zero variable cost assumptions.
6. The formulation of model "end effects".
7. The digit precision of the input parameters.

More broadly, we can conclude that model differences cannot be imprecisely attributed to an undefined "complexity." Because this work focuses on the intercomparison of highly simplified versions of four different models, it presents only a small sample of differences that can exist between full-scale capacity expansion models. The differences at each iteration and the active participation of model group members in resolving those differences only highlight the potential for significant differences among the more complex model formulations common in the research literature and planning studies. While more complex models will be undoubtedly more difficult to benchmark, we argue that doing so is important for a range of reasons discussed in this paper, including building model confidence and substantiating specific modeling choices. Model developers can tackle this formidable task by starting with unit tests that divide full model testing into smaller components. The simple cases and datasets presented in this study can be used as the foundation for development of additional tests to assess greater complexity.

It should be noted that while the results of the models studied here differed in detail, all of the models agreed on the same ranking of scenarios with respect to cost, with Wind Only (S4) as the highest cost, followed by Solar and Storage Only (S3), then Nuclear Only (S5), then Current Costs (S1), and Hypothetical Costs (S2) being the lowest cost. This gives us a sense of the resolution of conclusions that can be drawn from individual models without needing to worry about model details.

Consistency among models allows for more productive collaboration between macro-energy modelers and the broader energy modeling community, as well as the possibility of developing modules that can be incorporated into different models. As benchmarking and intercomparison efforts increase and we gain more understanding of the differences among models and why these differences occur, we can more effectively collaborate with and integrate our analyses with unit commitment, power flow, and integrated assessment models. This type of community-wide collaboration would enhance the existing work of many researchers who have been investigating linkages among capacity expansion models, unit commitment models, power flow models, integrated assessment models, and more. Integration of different types of energy models can help us to better represent the energy system as a whole and identify potential future development paths. In addition, collaboration within the modeling community could provide transparency regarding the energy transition and energy challenges, for policymakers and other stakeholders.

CRediT authorship contribution statement

Candise L. Henry: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Hadi Eshraghi:** Methodology, Formal analysis, Investigation, Writing – review & editing. **Oleg Lugovoy:** Methodology, Formal analysis, Investigation, Writing – review & editing, Funding acquisition. **Michael B. Waite:** Methodology, Formal analysis, Investigation, Writing – review & editing. **Joseph F. DeCarolis:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing, Funding acquisition. **David J. Farnham:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing. **Tyler H. Ruggles:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing. **Rebecca A.M. Peer:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing. **Yuezi Wu:** Methodology, Investigation. **Anderson de Queiroz:** Methodology, Formal analysis, Investigation. **Vladimir Potashnikov:** Methodology. **Vijay Modi:** Methodology, Formal analysis, Funding acquisition. **Ken Caldeira:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the Editor and two anonymous referees for many constructive comments on an earlier version of this work that have greatly improved its contents.

Appendix A

See Tables A1, B1 and B2.

Table A1
Cost and technology input values used to benchmark the four models.

	Solar	Wind	Natural Gas Combined Cycle	Nuclear	Battery
Cases S1, S3, S4, and S5					
Capital Cost ($/kW, or $/kWh marked with asterisk)	1851	1657	982	5946	261*
Capital Recover Factor (%/year)	8.06	8.06	9.44	7.50	14.24
Fixed O&M Cost ($/kW, or $/kWh marked with asterisk)	22.02	47.47	11.11	101.28	0*
Variable O&M Cost ($/kWh)	0	0	0.00354	0.00232	0
Fuel Cost ($/kWh)	0	0	0.0191	0.0075	–
Conversion Efficiency (%)	–	–	54	33	–
Charging Efficiency (%)	–	–	–	–	90
Decay Rate (fraction per hour)	–	–	–	–	1.14×10^{-6}
Charging Time (hours)	–	–	–	–	6.008
Project Life (years)	30	30	20	40	10
Discount Rate (% per year)	7	7	7	7	7
Case S2					
Capital Cost ($/kW, or $/kWh marked with asterisk)	788	1095	982	1027	26
Capital Recover Factor (%/year)	8.06	8.06	9.44	7.50	14.24
Fixed O&M Cost ($/kW, or $/kWh marked with asterisk)	22.02	47.47	11.11	101.28	0
Variable O&M Cost ($/kWh)	0	0	0.00354	0.00232	0
Fuel Cost ($/kWh)	0	0	0.0191	0.0075	–
Conversion Efficiency (%)	–	–	–	33	–
Charging Efficiency (%)	–	–	–	–	90
Decay Rate (fraction per hour)	–	–	–	–	1.14×10^{-6}
Charging Time (hours)	–	–	–	–	6.008
Project Life (years)	30	30	20	40	10
Discount Rate (% per year)	7	7	7	7	7

Table B1
The initial (iteration 1) outputs for each of the models in this analysis. These numbers are presented in Fig. 1A of the main text.

	Temoa					MEM					energyRt					SECTR				
Capacities (GW)	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Solar	0.0	113.3	4880.5	0.0	0.0	0.0	246.7	5039.1	0.0	0.0	0.0	213.7	5060.0	0.0	0.0	0.0	247.0	4879.8	0.0	0.0
Wind	0.0	0.0	0.0	11541.8	0.0	0.0	46.8	0.0	11541.3	0.0	0.0	34.9	0.0	11541.3	0.0	0.0	47.0	0.0	11541.3	0.0
NGCC	716.7	295.7	0.0	0.0	0.0	716.7	168.6	0.0	0.0	0.0	716.7	152.9	0.0	0.0	0.0	716.7	158.2	0.0	0.0	0.0
Nuclear	0.0	375.8	0.0	0.0	716.7	0.0	349.9	0.0	0.0	716.7	0.0	376.3	0.0	0.0	716.7	0.0	360.0	0.0	0.0	716.7
Battery (GWh)	*0.0*	*0.0*	*11717.3*	*0.0*	*0.0*	*0.0*	*857.4*	*10448.3*	*0.0*	*0.0*	*0.0*	*869.4*	*10557.0*	*0.0*	*0.0*	*0.0*	*903.6*	*12190.6*	*0.0*	*0.0*
Objective Function (B$)	237.8	215.3	970.2	2168.1	499.9	230.3	202.1	1253.5	2093.3	498.9	214.9	187.3	1168.6	1952.3	450.1	230.0	201.5	1288.6	2089.3	492.4

Table B2
The final (iteration 4) outputs for each of the models in this analysis. These numbers are presented in Fig. 1B of the main text.

	Temoa					MEM					energyRt					SECTR				
Capacities (GW)	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
Solar	0.0	246.6	4887.9	0.0	0.0	0.0	246.7	5039.1	0.0	0.0	0.0	246.7	5039.1	0.0	0.0	0.0	246.7	4887.9	0.0	0.0
Wind	0.0	46.7	0.0	11541.8	0.0	0.0	46.8	0.0	11541.3	0.0	0.0	46.8	0.0	11541.3	0.0	0.0	46.8	0.0	11541.3	0.0
NGCC	716.7	158.2	0.0	0.0	0.0	716.7	168.5	0.0	0.0	0.0	716.7	158.2	0.0	0.0	0.0	716.7	158.2	0.0	0.0	0.0
Nuclear	0.0	360.2	0.0	0.0	716.7	0.0	350.0	0.0	0.0	716.7	0.0	360.2	0.0	0.0	716.7	0.0	360.2	0.0	0.0	716.7
Battery (GWh)	*0.0*	*853.9*	*11533.9*	*0.0*	*0.0*	*0.0*	*857.4*	*10448.3*	*0.0*	*0.0*	*0.0*	*857.4*	*10448.3*	*0.0*	*0.0*	*0.0*	*903.8*	*12207.4*	*0.0*	*0.0*
Objective Function (B$)	229.9	201.3	1258.1	20,917	492.4	230.5	202.0	1253.5	2093.3	498.9	230.0	201.4	125.4	2088.9	492.4	230.0	201.5	1290.6	2089.3	492.4

References

[1] Blair N, Jenkin T, Milford J, Short W, Sullivan P, Evans D, et al. Renewable Energy and Efficiency Modeling Analysis Partnership (REMAP): An Analysis of How Different Energy Models Addressed a Common High Renewable Energy Penetration Scenario in 2025. NREL Technical Report, TP-6A2-45656; 2009.

[2] Mai T, Bistline J, Sun Y, Cole W, Marcy C, Namovicz C, et al. The role of input assumptions and model structures in projections of variable renewable energy: a multi-model perspective of the U.S. electricity system. Energy Econ 2018;76:313–24.

[3] Loulou R, Goldstein G, Noble K. Documentation for the MARKAL family of models. Energy Technology Systems Analysis Programme; 2004.

[4] Karlsson K, Meibom P. Optimal investment paths for future renewable based energy systems – Using the optimization model Balmorel. Int J Hydrogen Energy 2008;33(7):1777–87.

[5] DeCarolis J, Hunter K, Sreepathi S. The TEMOA Project: Tools for Energy Model Optimization and Analysis. International Energy Workshop; 2010.

[6] Howells M, Rogner H, Strachan N, Heaps C, Huntington H, Kypreos S, et al. OSeMOSYS: The Open Source Energy Modeling System: An introduction to its ethos, structure and development. Energy Policy 2011;39(10):5850–70.

[7] Short W, Sullivan P, Mai T, Mowers M, Uriarte C, Blair N, et al. Regional Energy Deployment System (ReEDS). National Renewable Energy Laboratory (NREL); 2011.

[8] Fripp M. Switch: a planning tool for power systems with large shares of intermittent renewable energy. Environ Sci Technol 2012;46(11):6371–8.

[9] Huber M, Roger A, Hamacher T. Optimizing long-term investments for sustainable development of the ASEAN power system. Energy 2015;88:180–93.

[10] Johnston J, Henriquez-Auba R, Maluenda B, Fripp M. Switch 2.0: A modern platform for planning high-renewable power systems. SoftwareX 2019;10.

[11] Lugovoy O, Potashnikov V. energyRt. Energy Systems Modeling R-Toolbox 2020. https://www.energyrt.org/.

[12] Dowling JA, Rinaldi KZ, Ruggles TH, Davis SJ, Yuan M, Tong F, et al. Role of long-duration energy storage in variable renewable electricity systems. Joule 2020;4(9):1907–28.

[13] Ruggles TH, Dowling JA, Lewis NS, Caldeira K. Opportunities for flexible electricity loads such as hydrogen production from curtailed generation. Adv Appl Energy 2021;3.

[14] Jenkins JD, Sepulveda NA. Enhanced Decision Support for a Changing Electricity Landscape: An MIT Energy Initiative Working Paper. MIT Energy Initiative, MITEI-WP-2017-10; 2017.

[15] Dyner Isaac, Larsen Erik R. From planning to strategy in the electricity industry. Energy Policy 2001;29(13):1145–54.

[16] Conejo Antonio J, Ruiz Carlos. Complementarity, not optimization, is the language of markets. IEEE Open Access J Power Energy 2020;7:344–53.

[17] Weyant John, Kriegler Elmar. Preface and introduction to EMF 27. Clim Change 2014;123(3-4):345–52.

[18] Edenhofer O, Lessmann K, Kemfert C, Grubb M, Kohlert J. Induced technological change: exploring its implications for the economics of atmospheric stabilization: synthesis report from the innovation modeling comparison project. Energy J 2006;2:57–108.

[19] Clarke L, Krey V, Weyant J, Chaturvedi V. Regional energy system variation in global models: results from the Asian Modeling Exercise scenarios. Energy Econ 2012;34(3):S293–305.

[20] Kriegler Elmar, Mouratiadou Ioanna, Luderer Gunnar, Bauer Nico, Brecha Robert J, Calvin Katherine, et al. Will economic growth and fossil fuel scarcity help or hinder climate stabilization? Overview of the RoSE multi-model study. Clim Change 2016;136(1):7–22.

[21] van der Zwaan BCC, Calvin KV, Clarke LE. Climate Mitigation in Latin America: Implications for Energy and Land Use: Preface to the Special Section on the findings of the CLIMACAP-LAMP project. Energy Econ 2016;56:495–8.

[22] Long JCS, Baik E, Jenkins JD, Kolster C, Chawla K, Olson A, et al. Clean firm power is the key to California's carbon-free energy future. Issues Sci Technol 2021.

[23] Koltsaklis Nikolaos E, Dagoumas Athanasios S. State-of-the-art generation expansion planning: a review. Appl Energy 2018;230:563–89.

[24] Morrison Robbie. Energy system modeling: public transparency, scientific reproducibility, and open development. Energy Strategy Rev 2018;20:49–63.

[25] Ringkjøb Hans-Kristian, Haugan Peter M, Solbrekke Ida Marie. A review of modelling tools for energy and electricity systems with large shares of variable renewables. Renew Sustain Energy Rev 2018;96:440–59.

[26] Subramanian Avinash, Gundersen Truls, Adams Thomas. Modeling and simulation of energy systems: a review. Processes 2018;6(12):238. https://doi.org/10.3390/pr6120238.

[27] Babatunde Olubayo M, Munda Josiah L, Hamam Yskandar. A comprehensive state-of-the-art survey on power generation expansion planning with intermittent renewable energy source and energy storage. Int J Energy Res 2019;43(12):6078–107.

[28] Groissböck Markus. Are open source energy system optimization tools mature enough for serious use? Renew Sustain Energy Rev 2019;102:234–48.

[29] Levi Patricia J, Kurland Simon Davidsson, Carbajales-Dale Michael, Weyant John P, Brandt Adam R, Benson Sally M. Macro-energy systems: toward a new discipline. Joule 2019;3(10):2282–6.

[30] IEEE Power Grid Lib. Institute of Electrical and Electronics Engineers Power and Energy Society Task Force on Benchmarks for Validation of Emerging Power System Algorithms, power-grid-lib.github.io.

[31] Yue Xiufeng, Pye Steve, DeCarolis Joseph, Li Francis GN, Rogan Fionn, Gallachóir Brian Ó. A review of approaches to uncertainty assessment in energy system optimization models. Energy Strategy Rev 2018;21:204–17.
[32] Ruggles TH, Farnham DJ, Tong D, Caldeira K. Developing reliable hourly electricity demand data through screening and imputation. Scientific Data 2020;7.
[33] Gelaro Ronald, McCarty Will, Suárez Max J, Todling Ricardo, Molod Andrea, Takacs Lawrence, et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J Climate 2017;30(14):5419–54.
[34] EIA (2018) Annual Energy Outlook 2018: with projections to 2050. U.S. Energy Information Administration. https://www.eia.gov/outlooks/archive/aeo18/pdf/AEO2018.pdf.
[35] Gurobi. Gurobi Optimization Documentation: Tolerances and user-scaling; 2021. https://www.gurobi.com/documentation/9.1/refman/tolerances_and_user_scalin.html.
[36] Eyring Veronika, Bony Sandrine, Meehl Gerald A, Senior Catherine A, Stevens Bjorn, Stouffer Ronald J, et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci Model Develop 2016;9(5):1937–58.
[37] American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). ANSI/ASHRAE Standard 140-2017 – Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs; 2017.