TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Doh, C., van Rijn, S., de Vries, L. J., & Cvetkovic, M. (2026). On the smart coordination of flexibility scheduling in multi-carrier integrated energy systems. *Scientific Reports*, *16*(1), Article 26927. https://doi.org/10.1038/s41598-026-57609-9

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

scientific reports

Check for updates

OPEN

On the smart coordination of flexibility scheduling in multi-carrier integrated energy systems

Christian Doh Dinga[1]✉, Sander van Rijn[2], Laurens J. de Vries[3] & Milos Cvetkovic[1]

Coordinating the interactions among flexibility assets in multi-carrier integrated energy systems (MIES) can lead to a cost-efficient energy transition. However, the proliferation of flexibility assets and their growing participation in active demand response increases the complexity of coordinating these interactions. This paper investigates several approaches to model the coordination of flexibility scheduling in MIES with many autonomous flexibility providers. We propose runtime model coupling as an alternative modeling paradigm to overcome the limitations of monolithic centralized co-optimization. Specifically, we introduced two model coupling approaches—a distributed price-response and a decentralized market auction approach—to address practical challenges such as preserving the autonomy and privacy of flexibility providers while ensuring scalability. We conduct a quantitative benchmark of these approaches against co-optimization across varying problem sizes, complexities, and computing infrastructures. This benchmark provides new empirical insights into the trade-offs between optimality, autonomy, and scalability that have so far remained unquantified in the energy system literature. We show that model coupling offers a method to balance optimality and realism (autonomy and privacy) while delivering substantial scalability gains. The proposed model coupling approaches are formalized as open-source software with several practical applications: modelers can experiment with different flexibility modeling approaches and choose the one that best matches their modeling objectives and constraints; flexibility providers can couple their models to simulate interactions between their systems to make informed operational decisions without disclosing any confidential information.

Keywords Energy system integration, Multi-energy systems, Distributed energy resources, Demand flexibility, Demand response, Electricity markets, Energy economics, Multi-agent simulation, Model coupling, Co-simulation, Distributed optimization

List of symbols

Sets and indices

$a \in A$	Set of prosumer assets
$b \in B$	Set of price-quantity bid blocks
$c \in C$	Set of energy carriers excluding electricity
$i \in N_g$	Set of generator agents
$j \in N_p$	Set of prosumer agents
$n \in N$	Set of all agents including generators and prosumers
$N_s^g \subseteq N_g$	Set of generator-satellites: generators represented as satellite agents that submit bids
$N_s^p \subseteq N_p$	Set of prosumer-satellites: prosumers represented as satellite agents that submit bids
$N_s = N_s^g \cup N_s^p$	Set of all satellite agents (both generator and prosumer satellites)
$N_c = N_g \setminus N_s^g$	Set of central-model generators: generators whose dispatch is determined directly by the central market-clearing problem
$t \in T$	Set of time steps

Parameters

α_i, β_i	Generator quadratic and linear cost coefficients ($20\ \mathrm{AC/MW^2h}$, $20\ \mathrm{AC/MWh}$)

[1]Department of Electrical Sustainable Energy, Delft University of Technology, 2628 CD Delft, The Netherlands. [2]Environment and Sustainability Section, Netherlands eScience Center, 1098 XH Amsterdam, The Netherlands. [3]Department of Engineering Systems and Services, Delft University of Technology, 2600 GA Delft, The Netherlands. ✉email: c.dohdinga@tudelft.nl

$\lambda_{j,t}^{forecast}$	Electricity price forecast (20 AC/MWh)
$\pi_{j,b,t}$	Price of bid block b (20 AC/MWh)
$\mu_{c,t}$	Cost of energy carrier (20 AC/MWh)
δ_a^p	Binary parameter: equals 1 if the conversion process at asset a involves electricity, 0 otherwise (–)
γ^q	Binary parameter: equals 1 if the conversion process involves heat, 0 otherwise (–)
$P_{j,t}^D$	Base demand of electricity (MW)
$R_{c,j,t}^D$	Base demand of non-electric carrier (MW)
$\eta_{a,j}^E$	Electrical efficiency of prosumer asset
$\eta_{a,j}^r$	Non-electric efficiency of prosumer asset
$\eta^{r,ch}$, $\eta^{r,dc}$	Storage charge/discharge efficiency
$\underline{I_{a,j}}$	Maximum capacity of converter asset (MW)
S_j^r	Maximum capacity of storage (MW)
$\underline{G_{i,t}}$, $\overline{G_{i,t}}$	Generator minimum, maximum available capacity (MW)
$\underline{E_j^{r,S}}$, $\overline{E_j^{r,S}}$	Storage minimum, maximum energy capacity (MWh)
$\underline{E_j^{t,F}}$, $\overline{E_j^{t,F}}$	Flexible demand min/max energy (MWh)
ρ	Penalty parameter in price-response coupling
k	Iteration number in price-response coupling
Δt	Time step length (h)
$\overline{\Phi}_{j,b,t}$	Maximum quantity submitted for prosumer bid block b (MW)
Primal and dual variables	
λ_t	Electricity (market clearing) price ($20 AC$/MWh)
$g_{i,t}$	Generator dispatch (MW)
$p_{j,t}$	Prosumer net electric power (MW)
$p_{gen}^{c,a,j,t}$	Prosumer electricity generated locally (MW)
$p_{a,j,t}^{con}$	Electricity consumed from the grid (MW)
$p_{a,j,t}^{ch}$, $p_{a,j,t}^{dc}$	(EV) battery charge/discharge (MW)
$q_{j,t}^{st}$	Heat generation from solar thermal (MW)
$r_{a,j,t}^{ch}$, $r_{a,j,t}^{dc}$	Non-electric storage charge/discharge (MW)
$p_{j,t}^F$, $r_{j,t}^F$	Flexible demand of energy carrier (MW)
$x_{c,a,j,t}$	Consumption rate of energy carrier (MW)
$r_{a,j,t}$	Generation rate of energy carrier (MW)
$e_{j,t}^{r,F}$	Energy level of flexible demand (MWh)
$e_{j,t}^{r,S}$	Energy level of (battery) storage (MWh)
$\phi_{j,b,t}$	Power quantity of bid block b (MW)

The growing electrification of energy demand and proliferation of distributed energy resources (e.g., heat pumps, electric vehicles, electrolyzers, etc.) increases the interactions and interdependencies between energy sectors and carriers[1–3], resulting in multi-carrier integrated energy systems (MIES). While energy system integration offers new sources of flexibility, the inclusion of many different types of distributed energy resources (herein, referred to as flexibility resources) increases the complexity of the energy system[4,5]. This is because it becomes challenging to coordinate interactions among demand-side flexibility resources[6]. Unlike traditional, centralized supply-side flexibility resources such as conventional generators that are (partially) controlled by the system operator, demand-side flexibility resources are controlled by prosumers who have full autonomy over their systems[7,8]. This implies that prosumers can directly, or through aggregators, participate in active demand response where they schedule their power consumption and/or generation predominantly based on price signals[9,10]. As shown in Harder et al.[11], excluding the autonomy of prosumers in energy system models can result in lower operational profits, sending misleading investment signals to flexibility providers, which are expected to play a crucial role in the energy transition[12–14]. Moreover, disregarding the autonomy of prosumers fails to capture the potential consequences of demand response. For example, recent studies have shown that uncoordinated response from a large pool of demand-side flexibility resources can aggravate congestion in the electricity grid if prosumers simultaneously ramp up consumption in response to low electricity prices[15,16]. Therefore, energy system integration brings along a methodological challenge—the challenge of coordinating the interactions among many different types of flexibility resources (owned and operated by prosumers with full autonomy) to optimally schedule flexibility.

In the literature, the conventional approach for modeling the coordination of flexibility scheduling (also referred to as demand response) in MIES is based on combined optimization (co-optimization). In this approach, all subsystems (prosumers or simply agents) are represented within a single, monolithic computational tool, and their interactions are coordinated by optimizing a global utility function, typically system-wide costs. Co-optimization has been widely applied in various MIES contexts. For instance, it has been used to coordinate flexibility scheduling in regional systems consisting of interconnected electricity, heat, and natural gas networks[17,18]. It has also been extended to account for uncertainty in the operation of MIES[19,20], as well as to assess the flexibility potential arising from increased electrification of heat demand in MIES[21,22]. Numerous other studies[23–25] adopt similar monolithic co-optimization approaches to model flexibility in MIES.

Although widely used, co-optimization has several drawbacks. First, its top-down control strategy assumes that agents give up control over their system, and that local control (dispatch) decisions are made centrally. This assumption takes away the autonomy of prosumers, which, as we already argued, is a key characteristic that reflects the operational realities of flexibility providers. Therefore, while co-optimization could be a valid approach for modeling a single agent's flexibility scheduling problem, it is, however, not suitable for studying

flexibility in a MIES where many agents with full autonomy interact. Second, co-optimization scales poorly with increasing problem size (e.g., number of agents), since all dispatch decisions are computed centrally (see Fig. 2). As the number of agents and the complexity of their models grow, the size of the resulting optimization problem increases significantly, which may lead to higher computational burden in terms of memory requirements and solution time. This challenge can be further exacerbated when long time horizons or intertemporal constraints—which are required to accurately represent flexibility—are considered. While advances in solver technology and high-performance computing (HPC) can alleviate some of these issues, it remains unclear to what extent co-optimization remains computationally tractable for large-scale MIES with many heterogeneous agents. Finally, perhaps, the most practical limitation of co-optimization is the need for information sharing, including confidential data. This is because to centrally compute dispatch decisions, co-optimization requires complete information of all agents (see Fig. 2), including the techno-economic parameters and internal states of their assets. This raises a privacy issue, as in reality, agents will not be willing to share such private information about their system for confidentiality reasons.

To address the aforementioned limitations of monolithic co-optimization in modeling the coordination of flexibility scheduling in large-scale MIES with many autonomous agents, we propose to use model coupling. Model coupling is a modeling paradigm that simulates an integrated system by combining independent, standalone models through the exchange of interface variables enabled by software techniques and algorithms[26,27]. The literature on model coupling spans a spectrum of strategies that differ fundamentally in when information is exchanged among coupled models. At one end of this spectrum are offline model coupling strategies, commonly known as soft-linking[28–32]. In offline or soft-linking coupling strategies, each model is executed in isolation without any interaction during runtime, and thus information exchange occurs only after each model completes its run; consequently, coupling is reduced to post-hoc data exchange. At the middle of this spectrum, are runtime model coupling strategies, also referred to as co-simulation[27,33]. Here, coupled models exchange interface variables during execution, while model interaction and time progression are managed by an explicit algorithmic coordination mechanism[27,33]. As a result, the computations (decisions) of each coupled model become mutually dependent and are updated dynamically throughout the simulation run. At the other far end of this spectrum are hard-linking strategies that integrate all submodels into a single monolithic model, eliminating inter-model interfaces entirely, thereby inheriting the properties and limitations of centralized co-optimization discussed above. Among these model coupling strategies, runtime model coupling is particularly well-suited for modeling flexibility in MIES with many autonomous agents. This is because offline coupling strategies cannot capture the dynamic, mutually dependent interactions that arise in flexibility scheduling, where the interactions among many models (agents) must be coordinated at runtime. Accordingly, in the remainder of this paper, we use the term model coupling to refer specifically to runtime coupling (co-simulation).

More broadly, our motivation for proposing model coupling for studying flexibility in MIES with many autonomous agents is threefold: autonomy (the locus of decision-making, that is, whether agents' dispatch decisions are computed locally or centrally, and whose utility that decision optimizes), privacy (the volume and granularity of techno-economic information that an agent must disclose to a coordinator or to other agents for the coordination to function), and scalability (how computational requirements, such as runtime and memory usage, grow with increasing problem size and complexity.) First, regarding autonomy, model coupling can enhance the autonomy of flexibility providers compared to monolithic co-optimization. This is because model coupling allows each agent to retain its own model and thus its decision-making-related computations, while interactions among agents are coordinated through information exchange protocols. As a result, agents can locally optimize their utility and compute their dispatch decisions, thereby enabling a more realistic representation of prosumer behavior and their interactions. Second, concerning privacy, because coupled models exchange only interface variables, private information such as techno-economic parameters or system internal states remains local to each model (agent), reducing the volume and granularity of information that prosumers must disclose regarding their systems. Finally, regarding scalability, the modularity of model coupling allows each agent (subsystem) to be simulated in its native tool using dedicated solvers; from a computational point of view, this modularity enables parallelism as independent models (agents) can be simulated concurrently, thereby promising substantial gains in computational efficiency and scalability.

We note that the autonomy and privacy advantages of model coupling are relative rather than absolute, and are best understood with respect to a baseline. Throughout this paper, the baseline is monolithic co-optimization. Accordingly, the assessments reported in Table 1 should be interpreted as qualitative comparisons

Criteria	Coordination approach		
	Co-optimization	**Price-response**	**Market-auction**
Modeling paradigm	Monolithic model	Model coupling	Model coupling
Coordination architecture	Centralized coordination	Distributed coordination	Decentralized coordination
Coordination mechanism	Centrally optimize a global utility through a one-shot optimization	Cooperatively optimize a global utility through iterations	Match agents' bids through rolling horizon-based market clearing
Required information exchange	Complete information of agents, including confidential parameters	Interface variables: dispatch and price time series, and system imbalance	Interface variables: demand and supply bid functions, and market price
Autonomy in agent's local decision-making	No autonomy: top-down control signals from a central optimizer	Partial autonomy: local optimization based on consensus price updates	Full autonomy: local optimization based on expected market price

Table 1. Taxonomy of various approaches for modeling the coordination of flexibility scheduling in MIES.

of how the three coordination approaches differ in the degree of autonomy retained by agents and the amount of information that must be disclosed to achieve coordination. A rigorous quantitative assessment of either concept is beyond the scope of the present work. For example, autonomy could be quantified through measures of the deviation between locally preferred and centrally imposed dispatch decisions, while privacy could be quantified using information-theoretic metrics of disclosure[34]. Developing and evaluating such metrics remains an important direction for future research. Moreover, increased privacy does not imply complete protection against the inference of information. As documented in optimization and control literature [34–36], the repeated exchange of dispatch and price signals in the price-response approach may, in principle, allow a coordinator to infer agent cost parameters through inverse-optimization techniques. Similarly, detailed price-quantity bids in the market auction approach may reveal information about an agent's underlying cost structure to a sufficiently informed market operator. Developing stronger guarantees against such inference attacks, for example, through differential privacy or related privacy-preserving mechanisms, also lies beyond the scope of this work and constitutes another promising direction for future research.

Model coupling has recently gained increasing attention in the energy systems literature as a means of enabling high-fidelity energy system assessments [32,37,38]. However, existing research on coupling energy models has focused primarily on the development of coupling infrastructures rather than on the coordination logic governing interactions between coupled models [39–41]. Consequently, the question of how to coordinate the interactions among operational models of different prosumers at runtime remains largely unexplored and poorly understood. This study, therefore, addresses this gap by investigating alternative strategies for coupling operational energy models of different types of flexible prosumers at runtime to effectively capture the dynamic, mutually dependent interactions required for coordinating flexibility scheduling in MIES. Depending on the required information exchange among the coupled models and on how their interactions are managed during runtime (i.e., the coordination mechanism), fundamentally different coordination architectures emerge, most notably, distributed and decentralized architectures. In this work, we introduce two model coupling approaches, each corresponding to one of the coordination architectures: a price-response approach based on a distributed coordination architecture, and a market auction approach based on a decentralized coordination architecture. Our architecture naming follows the convention in the energy system optimization literature[34,42] in which distributed requires a coordinator agent that broadcasts a signal (penalty or incentive) to steer the system towards equilibrium, and decentralized requires no such coordinator agent. It is worth noting that other alternative coordination paradigms exist for studying multi-agent energy systems, notably multi-agent reinforcement learning[43] and game-theoretic equilibrium models[44]. However, both lie outside the scope of this paper as our primary focus is on coordination paradigms based on model coupling.

In the proposed price-response model coupling approach (see Fig. 3), agents locally optimize their dispatch decisions and interact to cooperatively optimize a global utility. The required information exchange includes: (i) point-wise demand-supply time series computed locally by each agent as a response to price time series broadcast by a coordinator agent; and (ii) the system imbalance computed as the difference between total demand and supply. Given that agents' dispatch must satisfy the system-wide power balance constraint that couples their decisions, the broadcast prices and agents' dispatch decisions are iteratively updated based on the imbalance until a consensus price is reached at which the imbalance is below a specified threshold. Generally, provided an equilibrium exists, this iterative procedure converges to a solution (almost) equal to that of centralized monolithic co-optimization. However, although optimal, this approach also has some drawbacks. First, the autonomy of agents is somewhat partial since their dispatch decisions are penalized over iterations to align them towards cooperatively optimizing a global utility. Second, the iterative nature of the coordination mechanism can lead to high computational costs, since decisions are recomputed over the full simulation horizon at each iteration. As a result, the approach can scale poorly if applied to large-scale MIES with long time horizons. Moreover, convergence is not guaranteed in general, particularly when agents' optimization problems are non-convex.

The market auction approach—which uses a decentralized coordination architecture instead (see Fig. 4)—presents an alternative model coupling strategy that can overcome the limitations of both monolithic co-optimization and the price-response approach. The market auction approach is characterized by the following: (i) it allows for full autonomy in agents' local dispatch decisions in a more realistic way. This is because it relaxes the assumption of a central agent optimizing dispatch decisions in a top-down control manner, or a coordinator agent penalizing agents' local dispatch decisions over iterations to align them towards a global optimum. This, however, comes with a sacrifice in optimality compared to co-optimization and the price-response approach; (ii) the coordination mechanism here simulates market auctions by matching agents' bids through social welfare maximization, requiring that information is exchanged within the resolution and horizon of the market clearing. This promises significant gains in scalability since the full horizon is partitioned into smaller windows and simulated in a rolling manner; (iii) it only requires agents to share their demand and supply bid functions that internalize their techno-economic and intertemporal constraints. How they construct these bids is their confidential trade strategy encapsulated in their respective models. In doing so, agents encode all their system details in a bid function. This enhances their privacy (relative to co-optimization), as it avoids the need to share any further information other than bids; (iv) interestingly, exchanging bid functions rather than point-wise time series (as in the price-response approach) eliminates the need for iterations, which can be prone to non-convergence, especially in the presence of non-convexities.

The fundamental differences between the coordination approaches we have discussed so far are summarized in Table 1. As shown, these approaches differ mainly in their modeling paradigm (monolithic vs. model coupling), coordination architecture, coordination mechanism, required information exchange, and their representation of autonomy in agents' local decision-making.

In sum, the main contributions of this paper with respect to the scientific literature include:

1. We propose a model coupling paradigm for modeling the coordination of flexibility scheduling in MIES. Unlike traditional monolithic approaches, model coupling (i) allows for a more realistic representation of prosumer behavior since it retains their autonomy (ii) enhances privacy since coupled models only exchange interface variables, and (iii) shows promising improvement in scalability since each coupled model can be simulated in parallel in its native tool. We introduce two such model coupling approaches with fundamentally different coordination architectures: a distributed price-response approach and a decentralized market-auction approach.
2. We conduct a systematic benchmarking of monolithic co-optimization and the two proposed model coupling approaches across varying problem sizes (spanning 30–350 agents and simulation horizons from one day to one week), problem complexity (LP vs MILP), and computing infrastructures (laptop vs HPC) . This benchmarking provides new empirical evidence on the computational trade-offs between autonomy (accuracy), theoretical optimality, and scalability across different flexibility modeling approaches that have so far remained unquantified in the energy system modeling literature.
3. We formalize and implement our proposed model coupling approaches as new open-source computational frameworks[45,46] that are freely available for the scientific community to explore and experiment with different approaches for modeling flexibility scheduling in MIES. The code for running all experiments conducted in this paper is also made open source and archived on Zenodo[47] to ensure that the results presented here are reproducible by third parties.
4. From a practical perspective, this work provides guidelines for modeling flexibility in MIES. More specifically, it demonstrates how different modeling architectures can be used to balance realism (in terms of agents' autonomy and privacy), theoretical optimality, and computational tractability, thereby providing insights into which approach is most suitable depending on modeling objectives, system scale, and available computational resources.

The remainder of this paper is structured as follows. The methodology section presents a detailed description and mathematical formulations of the three coordination approaches for modeling flexibility scheduling in MIES. The experiment design section describes the experimental setup, case study, and data sources. The results and discussion section presents the findings. Finally, the conclusion section summarizes the key insights from this study.

Methodology

This section provides a description of the physical and economic interactions between flexibility assets and agents, and the problem formulations under the three coordination approaches. While any coordination objective could be used, we assume that the objective of each agent is to minimize costs or maximize revenues. The formulations presented here are restricted to LP and MILP problems for the sake of simplicity, and are based on the MIES shown in Figure 1, where each subsystem (enclosed in dashed rectangles) representing an agent is treated as a copper-plate following Brown et al.[48]. We further assume that all prosumers and generators are connected to a single electric bus. This modeling choice primarily aims to isolate the coordination layer,

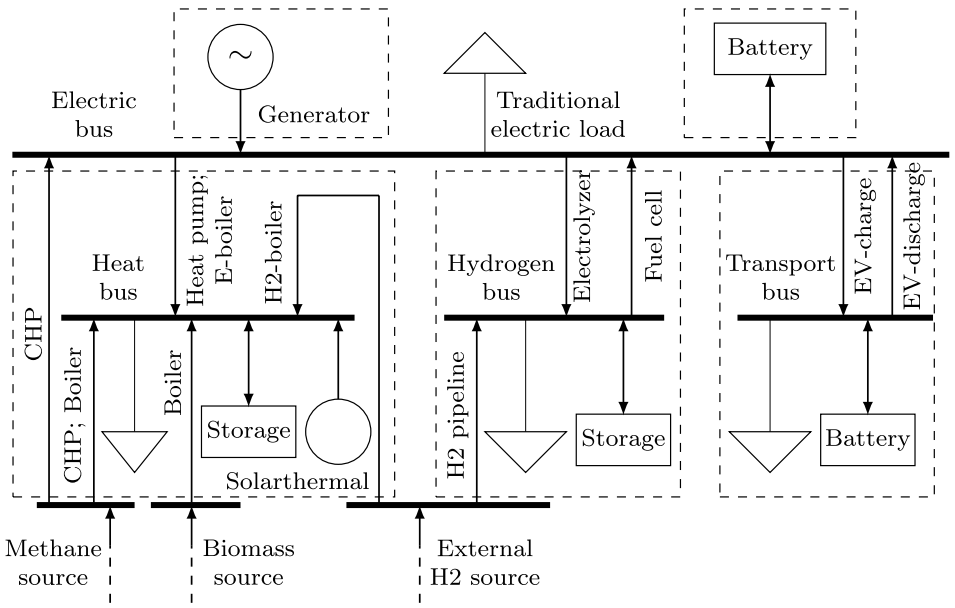


Fig. 1. Schematic of the multi-carrier integrated energy system (MIES) used as the reference for mathematical formulations throughout this paper. The electricity, heat, hydrogen, and transport subsystems are coupled through Power-to-X and X-to-Power conversion technologies. Each subsystem enclosed in a dashed rectangle represents an autonomous agent.

which is the central object of study in this work. Specifically, we aim to compare the scaling behavior of the three coordination paradigms under a common physical representation of the MIES. This abstraction is also consistent with the current European electricity market design, in which day-ahead market clearing is performed at the zonal resolution, while intra-zonal congestion is managed ex-post through redispatch[49]. Nevertheless, the formulations presented here can be easily extended to spatially-resolved MIES; see, for example, Neumann et al.[50]. The implications of relaxing the copper-plate assumption, and the extent to which the findings of this study transfer to spatially-resolved MIES, are discussed in detail in the "Results and discussion" section.

The MIES shown in Fig. 1 consists of heat, hydrogen, and transport systems coupled to the electricity system through Power-to-X and X-to-Power technologies. The electricity system has generators, batteries, and traditional loads that represent demand from passive consumers. Each physical system enclosed in a dashed rectangle represents an agent with full autonomy. Taking the heat subsystem as an example, the heat prosumer has a heat load that can be satisfied through multiple assets. For example, when electricity prices are lowest, the heat prosumer could choose to dispatch electric heating assets (e.g., heat pumps), shifting demand from other carriers to electricity to minimize costs. However, when electricity prices are highest, the prosumer can, for instance, dispatch its methane or biomass-based CHP asset to meet heat demand while also generating revenues from electricity sales. The heat prosumer also has a heat storage to buffer heat production and increase flexibility by decoupling demand and supply in time.

Co-optimization

Under co-optimization, the coordination problem is formalized as a centralized optimization problem where the dispatch of all assets is determined by minimizing the operational costs to satisfy the demand for all energy carriers. The coordination architecture that illustrates how the interactions in co-optimization would happen in real-world is depicted in Fig. 2.

First, generator and prosumer agents share all their local information (e.g., techno-economic parameters, internal states of their systems, loads, load shifting potential, etc.) with a central agent. The central agent then computes all dispatch decisions centrally and sends the optimal operational schedules to all agents to be executed in a top-down manner. Mathematically, the problem can be written as follows.

$$\min \quad \sum_{i,t}\left(\alpha_i \cdot g_{i,t}^2 + \beta_i \cdot g_{i,t}\right) + \sum_{c,a,j,t}\left(\mu_{c,t} \cdot x_{c,a,j,t}\right) \tag{1}$$

subject to:

$$\sum_{i=1}^{N_G} g_{i,t} + \sum_{j=1}^{N_P} p_{j,t} = 0 \qquad \forall t: \quad \lambda_t \tag{2}$$

$$p_{j,t} = \sum_{c,a} p_{c,a,j,t}^{gen} + \sum_{a} p_{a,j,t}^{dc} - \sum_{a}\left(p_{a,j,t}^{con} + p_{a,j,t}^{ch}\right) - \left(p_{j,t}^{F} + P_{j,t}^{D}\right) \qquad \forall j,t \tag{3}$$

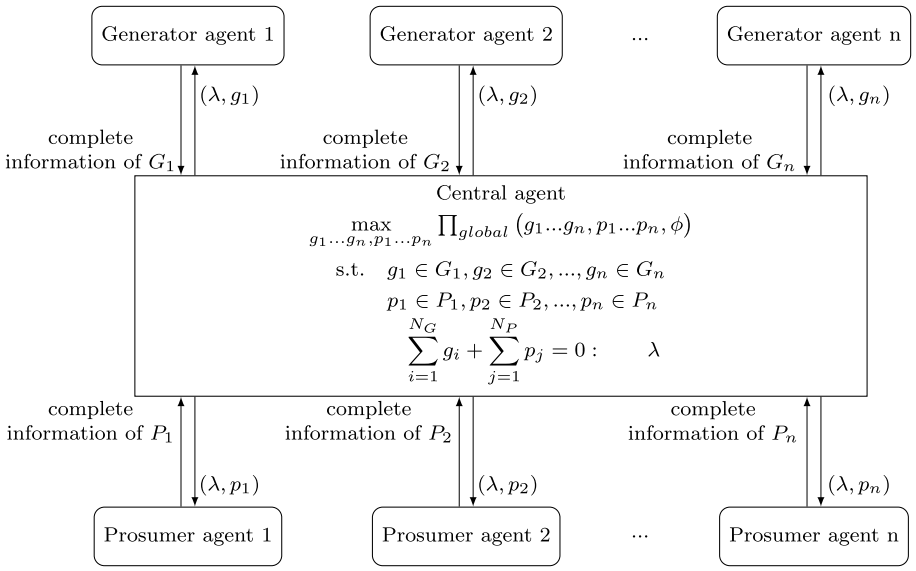


Fig. 2. Interactions under centralized monolithic co-optimization. All generator and prosumer agents disclose their complete techno-economic parameters and internal states to a central agent, who solves a single system-wide optimization problem. The resulting per-agent optimal dispatch is then sent to each agent to be executed in a top-down control manner.

$$p^{gen}_{c,a,j,t} = \eta^{E}_{a,j} \cdot x_{c,a,j,t} \qquad \forall c, a, j, t \tag{4}$$

$$r_{a,j,t} = \eta^{r}_{a,j} \cdot \left(\sum_{c \neq r} x_{c,a,j,t} + \delta^{p}_{a} \cdot p^{con}_{a,j,t} \right) \qquad \forall r \in \{q, h\}, a, j, t \tag{5}$$

$$\sum_{a} r_{a,j,t} + \gamma^{q} \cdot q^{st}_{j,t} + r^{dc}_{j,t} - r^{ch}_{j,t} = r^{F}_{j,t} + R^{D}_{c=r,j,t} \qquad \forall r \in \{q, h\}, j, t \tag{6}$$

$$e^{r,F}_{j,t} = e^{r,F}_{j,t-1} + r^{F}_{j,t} \cdot \Delta t \qquad \forall r \in \{p, q, h\}, j, t \tag{7}$$

$$e^{r,S}_{j,t} = e^{r,S}_{j,t-1} + \eta^{r,ch} \cdot r^{ch}_{j,t} \cdot \Delta t - \frac{r^{dc}_{j,t} \cdot \Delta t}{\eta^{r,dc}} \qquad \forall r \in \{p, q, h\}, j, t \tag{8}$$

$$\underline{G_{i,t}} \leq g_{i,t} \leq \overline{G_{i,t}} \qquad \forall i, t \tag{9}$$

$$0 \leq x_{c,a,j,t} \quad p^{con}_{a,j,t} \leq \overline{I_{a,j}} \qquad \forall c, a, j, t \tag{10}$$

$$0 \leq r^{ch}_{j,t} \quad r^{dc}_{j,t} \leq \overline{S^{r}_{j}} \qquad \forall r \in \{p, q, h\}, j, t \tag{11}$$

$$\underline{E^{r,S}_{j}} \leq e^{r,S}_{j,t} \leq \overline{E^{r,S}_{j}} \qquad \forall r \in \{p, q, h\}, j, t \tag{12}$$

$$\underline{E^{r,F}_{j}} \leq e^{r,F}_{j,t} \leq \overline{E^{r,F}_{j}} \qquad \forall r \in \{p, q, h\}, j, t \tag{13}$$

The objective function defined by Eq. (1) minimizes the cost of electricity generation and the consumption costs of externally priced energy carriers, such as methane and biomass. The end-use energy carriers: electricity, heat, and hydrogen, are denoted by p, q, h, respectively. Equation (2) ensures that electricity generation and consumption must be balanced at all time steps. The prosumer electric power defined in Eq. (3) is the difference between the electricity generated locally and that consumed from the grid or market; it is negative for net consumption and positive for net generation. Constraints defined by Eqs. (4) and 5 capture the conversions between energy carriers; the local balance of heat and hydrogen is enforced by Eqs. (6), (7) and 8 capture the energy dynamics of flexible demand and storage; Eqs. 9 to 11 limit the dispatch of generators and prosumer assets within their technical capacities; and finally, the constraints defined by Eqs. 12 and 13 model the minimum and maximum state-of-charge of storage and the feasible energy levels of flexible demands. We assume that flexible demand can be shifted in time but is not curtailable. We exclude load shedding from the scope of *flexibility* considered in this paper, as it represents a reduction in service rather than a temporal redistribution of energy use. Otherwise, modeling such behavior would require additional assumptions about willingness-to-pay or value-of-lost-load that are outside the scope of this work. The non-electric carrier (methane, biomass, and hydrogen) prices $\mu_{c,t}$ are deterministic, time-varying exogenous parameters and thus, are not formed endogenously by market clearing in the model. Only the electricity price λ_t—the dual variable of the electric power balance in Eq. 2, and thus the shadow price of electricity—is determined endogenously in the model. Furthermore, the objective in Equation (2) does not include a utility term for flexible demand, that is, load shifting is assumed to be free of inconvenience cost. Generators are also not subject to ramping constraints and thus can adjust intertemporal dispatch instantaneously. While these assumptions are consistent with our objective of isolating the effects of different coordination approaches from additional modeling complexities, such assumptions will tend to favor co-optimization; therefore, any optimality gap reported between co-optimization and the two proposed model coupling approaches should be interpreted as a lower bound.

Price-response model coupling

In the price-response approach, the coordination problem is formalized as a set of optimization problems, each representing the dispatch optimization of a generator or prosumer. A visualization of the interactions in this approach is illustrated in Fig. 3. As shown, it is an iterative process that starts with a coordinator agent broadcasting a price time series and system imbalance to agents. Given this information, agents run their local optimization to determine their optimal dispatch. Next, they report back their electric power generation or consumption time series to the coordinator agent, who then calculates the new system imbalance. If the stopping criteria (system imbalance defined by Eq. 16 and change in agents' decisions over iterations defined by Eq. 19) is not below the predefined threshold, a new price timeseries is computed and broadcast, and the process repeats again. The pseudo-code of the price-response coordination mechanism is shown in Algorithm 1. In what follows, we define the optimization problem of each agent type in iteration k.

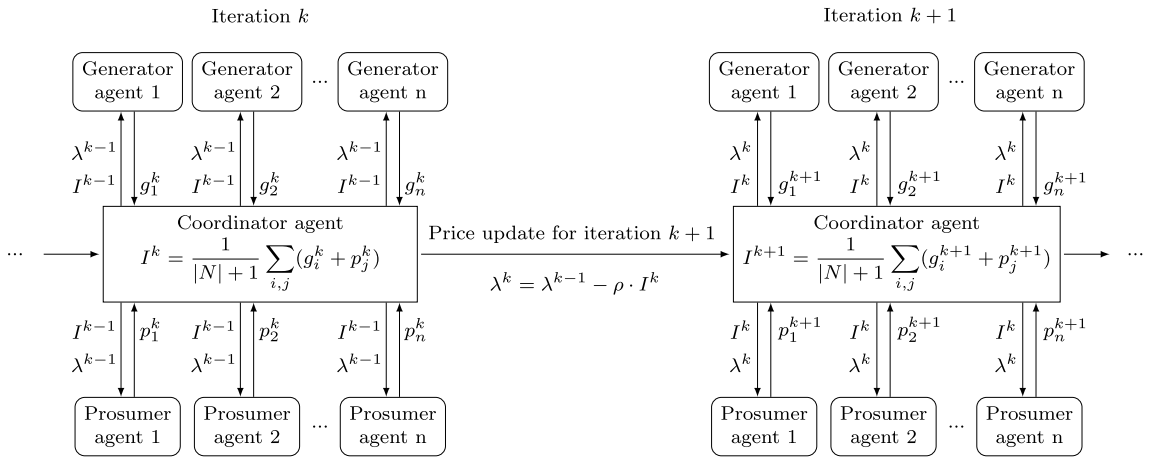


Fig. 3. Interactions under price-response model coupling. Each generator and prosumer agent solves its own local optimization problem and exchanges only its dispatch decision (techno-economic parameters and internal states remain local) and a consensus price signal with the coordinator agent, who iteratively updates this price signal until convergence.

Require: Initial price λ_t^0 for all t; initial penalty $\rho^0 > 0$; maximum iteration count $K_{\max}$; convergence tolerance ε; initial dispatch $g_{i,t}^0 = 0$ for all $i \in N_g$ and $p_{j,t}^0 = 0$ for all $j \in N_p$; initial imbalance $I_t^0 = 0$ for all t

$k \leftarrow 0$
while $k < K_{\max}$ **do**
  $k \leftarrow k + 1$
Step 1: Local agent updates (executed in parallel).
  for all $i \in N_g$ **in parallel do**
    Compute $g_{i,t}^k$ by solving Equation 14
  end for
  for all $j \in N_p$ **in parallel do**
    Compute $p_{j,t}^k$ by solving Equation 15
  end for
Step 2: Coordinator gathers responses and computes the system imbalance.
  $I_t^k \leftarrow \frac{1}{|N|+1} \cdot \left(\sum_{i \in N_g} g_{i,t}^k + \sum_{j \in N_p} p_{j,t}^k \right) \quad \forall t$ ▷ Eq. 16
Step 3: Primal stopping criterion.
  $\Psi^k \leftarrow \left\| I^k \right\|_2$ ▷ Eq. 18
Step 4: Dual stopping criterion (well-defined for $k \geq 2$).
  if $k \geq 2$ **then**
    Compute $\widetilde{\Psi}^k$ from Equation 19
  else
    $\widetilde{\Psi}^k \leftarrow +\infty$
  end if
Step 5: Price update and broadcast to agents.
  $\lambda_t^k \leftarrow \lambda_t^{k-1} - \rho^{k-1} \cdot I_t^k \quad \forall t$ ▷ Eq. 17
Step 6: Dynamic penalty-parameter update (residual balancing).
  if $\Psi^k > 10 \cdot \widetilde{\Psi}^k$ **then**
    $\rho^k \leftarrow 2 \cdot \rho^{k-1}$ ▷ Primal stagnant: amplify the penalty
  else if $\widetilde{\Psi}^k > 10 \cdot \Psi^k$ **then**
    $\rho^k \leftarrow \rho^{k-1}/2$ ▷ Dual stagnant: relax the penalty
  else
    $\rho^k \leftarrow \rho^{k-1}$ ▷ Residuals balanced: keep the penalty
  end if
Step 7: Joint convergence check.
  if $\Psi^k \leq \varepsilon$ **and** $\widetilde{\Psi}^k \leq \varepsilon$ **then**
    break
  end if
end while
return committed dispatch $\{g_{i,t}^k, p_{j,t}^k\}_{i,j,t}$, equilibrium prices $\{\lambda_t^k\}_t$, iteration count k, and residual history $\{\Psi^k, \widetilde{\Psi}^k\}_{k=1}^K$

Algorithm 1. Price-response coordination mechanism.

Generator agents
Each generator agent maximizes its revenues from electricity generation within its technical limits by solving the following optimization problem.

$$\left.\begin{aligned} \max \sum_t \left(\lambda_t^{k-1} - (\alpha_i \cdot g_{i,t}^k + \beta_i) \right) \cdot g_{i,t}^k - \frac{\rho}{2} \cdot \left\| g_{i,t}^k - \left(g_{i,t}^{k-1} - I_t^{k-1} \right) \right\|_2^2 \\ \text{subject to:} \quad \text{constraints defined by Equation 9} \end{aligned}\right\} \quad \forall i \in N_g \tag{14}$$

The first term in the objective represents revenue maximization, while the second term is an augmented penalty term. The penalty serves to dampen oscillations by constraining abrupt changes in agents' local dispatch decisions across iterations. This helps drive the iterative process towards an equilibrium state by penalizing agents' dispatch decisions based on the system imbalance. This enforces agents to align their local decisions towards a global optimum while still allowing them to optimize locally, resulting in partial autonomy in their decision-making. Although the penalty is applied across all time steps, it acts only as an iterative proximal

regularizer that vanishes at equilibrium; it is zero at equilibrium. Consequently, the penalty term does not affect the dispatch of agents at the equilibrium solution; see Boyd et al. [51] for more details on this algorithm.

Prosumer agents

Each prosumer agent maximizes revenues from electricity generated locally, or minimizes the consumption costs of energy carriers, subject to its technical constraints. The following problem is established for each prosumer.

$$\left.\begin{array}{l} \min \sum_{t}\left(-\lambda_{t}^{k-1} \cdot p_{j,t}^{k}\right)+\sum_{c,a,t}\left(\mu_{c,t} \cdot x_{c,a,j,t}\right)+\frac{\rho}{2} \cdot\left\|p_{j,t}^{k}-\left(p_{j,t}^{k-1}-I_{t}^{k-1}\right)\right\|_{2}^{2} \\ \text{subject to:} \quad \text{constraints defined by} \quad \text{Equations } 3-8 \quad \text{and} \quad \text{Equations } 10-13 \end{array}\right\} \quad \forall j \in N_{p} \quad (15)$$

As mentioned earlier, when net consuming, $p_{j,t}^{k}$ is negative, and the term related to electricity in the objective becomes cost minimization; when net generating, $p_{j,t}^{k}$ is positive, and the electricity-related term becomes revenue maximization.

The coordinator agent

The role of the coordinator is to calculate the imbalance of the system in each iteration as in Eq. 16, and to update prices for the next iteration according to Eq. 17.

$$I_{t}^{k}=\frac{1}{|N|+1} \cdot\left(\sum_{i} g_{i,t}^{k}+\sum_{j} p_{j,t}^{k}\right) \quad \forall t \quad (16)$$

$$\lambda_{t}^{k+1}=\lambda_{t}^{k}-\rho \cdot I_{t}^{k} \quad (17)$$

The price update is done in the direction that reduces the imbalance: for example, if there is excess demand in iteration k, the price in iteration $k+1$ is increased, and vice versa. The imbalance is normalized by the number of subproblems contributing primal updates: $|N|+1$ represents generator and prosumer agents plus the coordinator agent that performs the price update.

The iterative process is controlled by two stopping criteria: a primal stopping criterion Ψ^{k} and a dual stopping criterion $\widetilde{\Psi}^{k}$. The primal residual, defined per time step, measures the violation of the power balance constraint. The primal residual stopping criteria at iteration k is therefore defined as the ℓ_2-norm of the imbalance residual vector $I^{k}=(I_{1}^{k}, \ldots, I_{|T|}^{k})$ as in Eq. (18).

$$\Psi^{k} \;=\; \left\|I^{k}\right\|_{2} \;=\; \sqrt{\sum_{t}\left(I_{t}^{k}\right)^{2}}. \quad (18)$$

The dual residual, defined per agent and per time step, measures the change in agents' decisions between consecutive iterations. It is multiplied by the penalty factor ρ to link the change in decisions to the change in prices. Similarly, the dual residual stopping criteria is defined as the ℓ_2-norm of the dual residual as in Eq. (19).

$$\widetilde{\Psi}^{k} \;=\left\|\rho \cdot\left[\left(g_{i,t}^{k}-I_{t}^{k}\right)-\left(g_{i,t}^{k-1}-I_{t}^{k-1}\right)\right]\right\|_{2} \;+\; \left\|\rho \cdot\left[\left(p_{j,t}^{k}-I_{t}^{k}\right)-\left(p_{j,t}^{k-1}-I_{t}^{k-1}\right)\right]\right\|_{2}. \quad (19)$$

The algorithm terminates when $\Psi^{k} \leq \epsilon$ and $\widetilde{\Psi}^{k} \leq \epsilon$ are simultaneously satisfied. In the standard ADMM formulation presented by Boyd et al. [51], separate tolerances are typically used for the primal and dual residuals, since they are not dimensionally equivalent and may differ in scale. Here, however, we adopt a single tolerance threshold ϵ for both Ψ^{k} and $\widetilde{\Psi}^{k}$ as a practical convergence criterion. This is made possible by using a dynamic penalty update rule, which dynamically adjusts ρ to keep the primal and dual residuals at comparable magnitudes during the iterative process. The dynamic penalty update strategy is often used to accelerate convergence speed[52,53].

The dynamic penalty parameter update rule

The penalty parameter ρ in Eqs. 14 and 15 plays a critical role in the convergence speed of the iterative procedure. A value of ρ that is too small leaves agents' local dispatch decisions only weakly coupled to the coordinator's price signal (slow primal convergence), whereas a value of ρ that is too large over-penalizes deviations and forces premature consensus before the coordinator has had a chance to refine the price (slow dual convergence). The classical residual-balancing strategy of Boyd et al. [51] addresses this by adjusting ρ between iterations in response to the relative magnitudes of the primal and dual residuals (see Step 6 of Algorithm 1): doubling ρ whenever the primal stopping criterion Ψ^{k} exceeds ten times the dual stopping criterion $\widetilde{\Psi}^{k}$, halving ρ whenever $\widetilde{\Psi}^{k}$ exceeds ten times Ψ^{k}, and otherwise keeping ρ unchanged. We adopt this rule in Step 6 of Algorithm 1.

In our empirical experiments, the dynamic rule materially increases the fraction of convex problem instances that converge within the iteration limit and reduces the mean iteration count when compared to a static-ρ. We emphasize that improving the underlying ADMM algorithm is not the focus of this paper: we adopt the dynamic-ρ update rule as a well-known modification that improves convergence[52,53], and otherwise treat the algorithm as a fixed building block of the price-response coordination architecture. Readers interested in a deeper treatment of penalty parameter selection and convergence acceleration techniques for ADMM are referred to Boyd et al. [51] and the surrounding literature.

The price-response model coupling software

We implement the price-response model coupling approach as a new open-source Python-based software package called Demoses-admm[45]. Demoses-admm provides a streamlined API for the exchange of messages—formatted as one-dimensional arrays—among the coupled models (agents). Inter-model communication is fully synchronous: within every iteration, the coordinator blocks until every agent has returned its updated dispatch before updating and proceeding to the next iteration. The software leverages the consensus form of ADMM, with a residual-balancing rule for the penalty parameter ρ, which is a well-known modification of the standard ADMM algorithm[51]. The implementation is structured in a modular manner, allowing individual models to be defined independently while remaining fully compatible with the communication protocol. This enables the flexible integration of an arbitrary number of heterogeneous models, as well as parallel execution of agent optimizations within each iteration. Demoses-admm abstracts key functionalities such as price update mechanisms, penalty parameterization, iteration control, message passing, and convergence monitoring. This design makes the framework highly customizable, allowing users to tune coordination settings while focusing on higher-level tasks, such as the formulation of agent-specific models, rather than the underlying coordination logic and algorithmic overhead.

Market auction model coupling

In the market auction approach, agents interact through a decentralized market mechanism, where dispatch and prices are determined through auctions as illustrated in Fig. 4. First, agents run their local optimization based on electricity price forecast scenarios to determine their demand and or supply bids. These bids, which reflect their willingness to generate or consume electricity at various price levels, are then submitted to the market operator agent. Next, the market operator clears the market, and the resulting market-clearing prices and dispatch quantities are communicated back to agents. In this setup, we refer to the market operator model as the central model and the agents' models as satellite models. The central market operator model and the agents' satellite models are described below, including their mathematical formulations.

The central model

The central model is an electricity market-clearing model that matches agents' demand and supply bids with the goal of maximizing social welfare. Within the framework, generators and prosumers can be represented as satellites that submit their own bids to the central model, or alternatively folded directly into the central market-clearing problem. Let $N_s = N_s^g \cup N_s^p$ denote the set of satellite agents, partitioned into generator-satellites $N_s^g \subseteq N_g$ (those whose bids are computed by their local optimization as in Eq. (23)) and prosumer-satellites $N_s^p \subseteq N_p$. Let $N_c = N_g \setminus N_s^g$ denote the set of generators whose dispatch is determined directly inside the central market-clearing problem; for these generators, their constraints can be directly included in the central model. The specific partition between N_c and N_s^g used in our experiments is reported in the "Experiment design" section. Under this generic formulation, the central market-clearing problem can be written mathematically as follows.

$$\begin{aligned} \max \quad & \sum_{j \in N_s, b, t} \left(\pi_{j,b,t} \cdot \phi_{j,b,t} \right) - \sum_{i \in N_c, t} \left(\alpha_i \cdot g_{i,t}^2 + \beta_i \cdot g_{i,t} \right) \\ \text{subject to:} \quad & \text{constraints defined by Equation 9 for } i \in N_c \end{aligned} \tag{20}$$

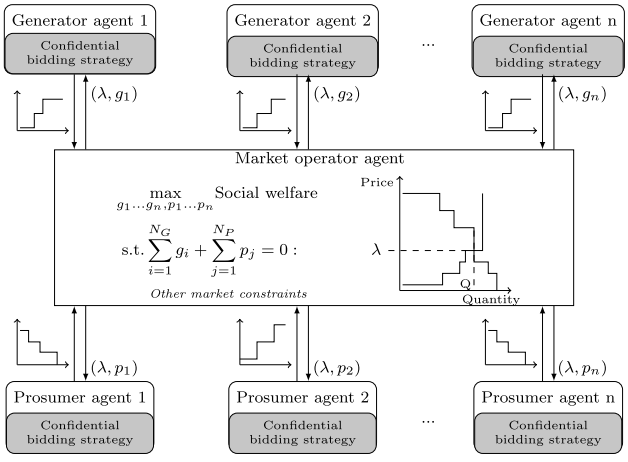


Fig. 4. Interactions under market auction model coupling. Each generator and prosumer agent constructs its demand/supply bids based on a local price forecast and submits them to the market operator; techno-economic parameters and internal states remain local. The market operator then clears the market and returns the clearing price and accepted quantities to each agent.

$$\sum_{i \in N_c} g_{i,t} + \sum_{j \in N_s} \sum_{b=1}^{B} \phi_{j,b,t} = 0 \quad \forall \quad t: \quad \lambda_t \tag{21}$$

$$0 \leq \phi_{j,b,t} \leq \overline{\Phi}_{j,b,t} \qquad \forall j \in N_s, b \in B, t \tag{22}$$

The objective defined in Eq. (20) maximizes social welfare defined as gross surplus from satellite bid blocks minus the generation cost incurred by central model generators. Equation (21) represents the market balance constraint, enforcing that the dispatch of central-model generators ($\sum_{i \in N_c} g_{i,t}$) plus the cleared bid blocks from satellite agents ($\sum_{j \in N_s, b} \phi_{j,b,t}$) sum to zero at every time step. Equation (22) caps the cleared quantity of each bid block to the maximum quantity $\overline{\Phi}_{j,b,t}$ submitted by satellite agent *j* for block *b* at time *t* to prevent the market from clearing unbounded quantities. By convention, a positive value of $\phi_{j,b,t}$ denotes a supply bid (net injection by satellite agent *j* into the electric bus) and a negative value denotes a demand bid (net withdrawal); both directions enter the market balance (Eq. 21) with a consistent sign so that supply bids are matched against demand bids and the residual is balanced by central model generators.

Satellite models

Each satellite agent runs a local optimization problem that maximizes its own objective subject to its local technical constraints. The form of this local problem differs between generator-satellites and prosumer-satellites.

For each generator-satellite $i \in N_s^g$, given an electricity price forecast $\lambda_{s,t}^{forecast}$ in scenario *s*, the local optimization problem is:

$$\left.\begin{array}{l} \max \sum_t \left(\lambda_{s,t}^{forecast} - (\alpha_i \cdot g_{i,t} + \beta_i)\right) \cdot g_{i,t} \\ \text{subject to:} \quad \text{constraints defined by} \quad \text{Equation 9} \end{array}\right\} \quad \forall s, i \in N_s^g \tag{23}$$

Since the focus of our study is on demand-side flexibility, we assume a simple bidding strategy for a generator-satellite: it submits a single bid block per time step at its marginal cost, with bid quantity equal to the optimal dispatch $g_{i,t}^*$ obtained from Eq. 23. This is functionally identical to the bidding behavior assumed for central model generators (where the bid takes the form of the quadratic supply bid term in Eq. 20); the difference is purely architectural—the dispatch is computed inside the satellite's local optimization rather than inside the central market-clearing problem—and serves to illustrate that the framework supports either modeling choice for any generator. For each prosumer-satellite $j \in N_s^p$, given an electricity price forecast $\lambda_{s,t}^{forecast}$ in scenario *s*, the local optimization problem is defined mathematically as follows:

$$\left.\begin{array}{l} \min \sum_t \left(-\lambda_{s,t}^{forecast} \cdot p_{j,t}\right) + \sum_{c,a,t} \left(\mu_{c,t} \cdot x_{c,a,j,t}\right) \\ \text{subject to:} \quad \text{constraints} \quad 3-8, \quad 10-13 \end{array}\right\} \quad \forall j \in N_s^p, \text{ for each forecast scenario } s \tag{24}$$

Satellite models implicitly account for uncertainty in market prices by generating bids over a range of different price forecast scenarios *s* rather than a single price-quantity bid pair. This provides some degree of robustness against deviations in forecasts, since bids are not tied to a single deterministic trajectory, but span a range of possible price realizations.

In the rolling-horizon implementation, each satellite agent's local optimization problem is solved repeatedly for consecutive market windows *w*. Within each window *w*, the time index *t* spans the look-ahead horizon $\mathcal{T}^w = \{t_w, t_w + 1, \ldots, t_w + L - 1\}$, where t_w is the start of the window and *L* is the length of the look-ahead window. Satellites with storage and/or flexible demand manage their intertemporal constraints in a look-ahead rolling-horizon manner through handing off states across windows by setting $e_{j,t_w-1}^{r,*} = e_{j,t_{w-1}^{\text{commit}}}^{r,*}$ for $r \in \{p, q, h\}$ and $* \in \{S, F\}$, where t_{w-1}^{commit} denotes the final committed time step of the previous market-clearing window. At each rolling-horizon step, only the dispatch corresponding to the committed portion of the current window is fixed, while the remaining future dispatch decisions are recomputed in the subsequent window. Satellite models can run with a look-ahead window of arbitrary length to optimize for multi-day scheduling. For a given price forecast, the result of the above prosumer-satellite optimization is the power dispatch time series $p_{j,t}$ and not the demand or supply bid functions required by the central model. To generate bid functions consisting of price-quantity pairs as required by the information exchange protocol in this approach, we developed a heuristic optimizer bidding algorithm for prosumer-satellites. The pseudo-code for the optimizer bidding algorithm developed for prosumer-satellites is shown in Algorithm 2.

Require: Prosumer j, time step t, CeilingPrice
1: $\Pi_j \leftarrow \{\text{CeilingPrice}\}$ ▷ Initialize Bid Prices
2: $\Pi_j \cup \{\min(\lambda_t^{forecast}, \ldots, \lambda_{t+f}^{forecast}) \forall f \in \text{Flexibilities}\}$
3: $\Pi_j \cup \{\mu_{c,t} \cdot \eta_{a,j}^r \forall c \in C\}$ ▷ Opportunity costs
4: **for all** $\pi_{j,b,t} \in \Pi_j$ **do**
5: $\lambda_t^{forecast} \leftarrow \pi_{j,b,t}$
6: $p'_{j,t} \leftarrow \text{simulateProsumerWithForecast}(\lambda^{forecast})$
7: $\phi_{j,b,t} \leftarrow p'_{j,t}$ ▷ bid = $(\pi_{j,b,t}, \phi_{j,b,t})$
8: **end for**
9: submitBids(B_j)

Algorithm 2. Bid generation algorithm for a prosumer agent.

In lines 1–3, all relevant bid prices are gathered: ceiling price, minimum forecast price per flexibility look-ahead range, and all opportunity costs for using different energy carriers. The "Flexibilities" on line 2 of Algorithm 2 is the set of shifting horizons attached to the prosumer's flexible demand categories, augmented with the full rolling-horizon look-ahead L as an additional entry—i.e., Flexibilities $= \{N : \text{the prosumer has a flexible-demand category with shifting horizon } N\} \cup \{L\}$. For each $f \in$ Flexibilities, the minimum forecast price over the interval $[t,\ t+f]$ is added as a candidate bid price. For example, a prosumer with two flexible demand categories of shifting horizons 4 h and 12 h would have Flexibilities $= \{4, 12, 24\}$ (with the third entry being the 24 h rolling-horizon look-ahead), and line 2 of Algorithm 2 would contribute three candidate bid prices: the minimum forecast price over $[t,\ t+4]$, over $[t,\ t+12]$, and over $[t,\ t+24]$. Once gathered, lines 4–8 determine the prosumer's demand. The forecast price for time t is set to the considered bid price (line 5), the prosumer reports their net electric power demand given that forecast (line 6), which is then set as the quantity of the associated bid (line 7). When that is done for every bid price, the collection of bids can be submitted as a bid function (line 9). To ensure that the submitted bid function is monotonic in price (as required by the central market-clearing model), we sort the resulting (π, ϕ) pairs by price and merge any non-monotonic segments into a single block (at the higher of the two prices) before submission.

The bid functions generated by Algorithm 2 are feasible by construction through four complementary mechanisms: (i) ceiling-price anchor: the bid-price set Π_j always contains the ceiling price (line 1), so the cumulative dispatch evaluated at the ceiling corresponds to demand that is guaranteed to clear; (ii) deadline-forced demand: flexible demand is internally modeled in the satellite optimization as a "store" with a lower-bound trajectory, computed from its shifting horizon and enforced through Eqs. 7 and 13. As the delivery deadline approaches, any residual flexible demand is therefore forced into the dispatch and bid at ceiling price; (iii) technically-feasible dispatch: every per-price dispatch query on line 6 is obtained from the satellite's LP/MILP with all technical constraints active, so each submitted (π, ϕ) bid pair corresponds to a feasible local dispatch; and (iv) market-side cap: Eq. 22 limits the cleared quantity of each bid block to the quantity offered by the prosumer, preventing the market from clearing more power than was submitted. After market clearing, the satellite reconciles its internal state through a deterministic re-solve with the cleared electric power and electricity price fixed.

The bidding strategy in Algorithm 2 assumes price-taking (non-strategic) behavior at the satellite level: each prosumer responds to the forecast price as an exogenous input and does not internalize its own market impact. Consequently, strategic bid manipulation and market-power effects are not considered. If market power were the question of interest, Algorithm 2 could be replaced by a strategic bidding formulation (e.g., a game-theoretic bi-level optimization) and extended to account for uncertainty through stochastic optimization. However, developing advanced "perfect" bidding strategies is beyond the scope of this work. Our rationale is that bidding strategies are inherently private to the agents, and thus it is the responsibility of each agent to construct their bids so that their privacy is enhanced. The present implementation in Algorithm 2, therefore, serves only as a proof-of-concept. Nevertheless, the modularity provided by the model coupling framework allows different bidding strategies to be implemented and evaluated with minimal changes to the coordination architecture and mechanism.

The market auction model coupling software

The market auction model coupling approach is implemented as a new open-source Python-based software Package called Annular [46]. Annular is used to couple the central market model to satellite models. The algorithms for each model are implemented generically using Cronian[54]—a companion package to Annular that builds optimization problems from a collection of `yaml` and `csv` files for model configuration and data. Annular provides a clear API for the communication and exchange of messages (bids) defined as multi-index tables. Therefore, the central and satellite models only have to be defined in terms of how they should interact (their conduits). Annular takes care of all overhead, such as communication and parallelization tasks, while letting each satellite model contain its own (private) bidding strategy. Similar to Demoses-admm[45], inter-model communication in Annular is also fully synchronous: the central model waits for bids from all satellites before performing market clearing, and subsequently broadcasts the cleared dispatch to all satellites prior to advancing

the rolling-horizon window. Therefore, the central model does not perform clearing on incomplete or stale information, and satellites only update their internal state after market clearing has been completed. Annular provides synchronized execution of, and communication between models, whether run on a laptop or a cluster system. Additionally, models in Annular can be pointed to different implementations or modeling tools and/or different bidding strategies, making Annular highly flexible and easily extensible.

Experiment design

Experiment setup

To evaluate the viability of our proposed model coupling approaches, we benchmark them against co-optimization by comparing agents' dispatch schedules and system costs. To investigate computational efficiency and scalability, we compare the absolute runtime and scaling performance, as well as the peak memory usage, for each approach. This benchmark provides insights into the trade-off between accuracy (preserving agents' autonomy), theoretical optimality, and scalability. We conduct multiple experiments to investigate computational performance by varying the following parameters.

Varying problem size

Problem size can be varied along two dimensions: time (simulation resolution or horizon) and space (number of agents). As a first experiment, we vary problem size while maintaining the same computing infrastructure. We increase problem size by increasing the simulation horizon from 1 day to a week (with hourly resolution) and the number of agents (coupled models) from 30 to about 350. In the temporal dimension, we focus on increasing the simulation horizon. While temporal resolution and horizon represent different modeling choices, both directly affect the total number of time steps and, consequently, the size of the resulting problem. Increasing the horizon, therefore, provides a representative way to capture the growth in computational complexity that would also arise from finer temporal resolutions, as both lead to larger decision spaces and stronger intertemporal coupling. It is also worth noting that while experiments with larger problem sizes could be performed, the selected problem sizes are sufficient to capture scaling trends along temporal and spatial dimensions without exhausting the available computational resources.

Varying problem complexity

As a second experiment, we repeat all problem instances from the first experiment with an added complexity by introducing non-convexity through binary decision variables. Specifically, we explicitly prevent the simultaneous charge and discharge of storage assets that might occur when prices are negative. In addition to Equations 8, 12, and 11 that model storage, the following constraints are added to the prosumer subproblem:

$$0 \leq r^{ch}_{j,t} \leq \omega_{j,t} \cdot \overline{S^r_j}, \qquad \forall j, t \tag{25}$$

$$0 \leq r^{dc}_{j,t} \leq (1 - \omega_{j,t}) \cdot \overline{S^r_j}, \qquad \forall j, t \tag{26}$$

$$\omega_{j,t} \in \{0, 1\}, \qquad \forall j, t \tag{27}$$

The binary variable $\omega_{j,t}$ takes value 1 when the storage at prosumer j is charging at time t and 0 otherwise, so that Eqs. 25 and 26 together enforce that the storage cannot simultaneously charge and discharge. We adopt storage complementarity as a representative non-convexity (other forms of non-convexity, such as unit commitment binaries and min-up/min-down constraints, can also be introduced) because it provides a minimal yet sufficient test of how each coordination approach behaves once the problem departs from the LP regime.

Varying computing infrastructure

The different modeling approaches make use of different coordination architectures, meaning they also differ in how much they can benefit from parallelism in the problem definition. Knowing how much a modeling approach can benefit from parallelism will help inform which trade-off the user or modeler can afford to make. If, for instance, a better time resolution can be achieved by using additional available computing power, that shifts the trade-off from time vs space to time vs computation cost. Naturally, there will also be a limit to how much benefit can be obtained from parallelism, beyond which it will not make sense to deploy the modeling approach on a system with more parallel CPU cores. To give an initial indication of this trade-off, we vary the computational resources used to run the benchmark between the authors' modern laptop with 16 CPU cores, and a more capable 48-core and 128-core server machine— Snellius[55], as a third experiment. This will indicate some actual possible speed-up from parallelism in the different modeling approaches.

Case study and data source

To use real energy system data for our experiments, we ran PyPSA-Eur model[50] with Germany only for the weather year 2013. The weather year 2013 is widely used as a representative reference year in European energy system studies, and is therefore adopted here for consistency with existing energy system modeling literature[50,56–59]. The collected data include asset capacities (generators, heat pumps, etc.), generator marginal costs, and hourly availability profiles for renewables; hourly demand profiles for electricity, heat, hydrogen, and transport; hourly price profiles for energy carriers such as methane, hydrogen, and biomass. We slightly increase the capacity of storage to reflect a future system with a substantial amount of flexibility.

A typical single node in a PyPSA-Eur model looks like the MIES in Fig. 1, and several of such nodes are interconnected through electricity transmission lines to form a network. Consistent with the copper-plate assumption introduced in the "Methodology" section—adopted to isolate the coordination layer and reflect

zonal electricity market pricing—we wrote Python scripts to flatten the original network, resulting in a copper-plate model in which all prosumers and generators are connected to a single electric bus. A prosumer is assumed to own all assets that feed into the bus to which it is attached. If there is more than one demand on a bus, the first is assumed to be the base demand (inflexible), while the others are assumed to be flexible. To vary the number of agents for our scaling experiments, we use PyPSA-Eur clustering functionality to increase the spatial granularity of the network to generate the required number of agents. The smallest network instance (with a spatial granularity of 1) results in about 30 agents, of which 16 are generators and 14 are prosumers. Increasing the spatial granularity by a step size of 1 increases the number of agents by roughly 10, and the proportion of distribution between agent types is constant; for every 10 agents added, about 6 are generators and 4 are prosumers.

The partition between the central model and satellite agents in the market auction experiments is defined as follows: all prosumers are modeled as satellites ($N_s^p = N_p$) for all problem instances; $|N_g| - 1$ generators are included in the central model while the remaining generator is modeled as a satellite ($N_s^g = 1$), submitting a single bid corresponding to its available capacity at marginal cost via the generator-satellite formulation in Equation 23. The reason for such a partition is twofold: first, it ensures that the focus of the study remains on demand-side flexibility, which is the primary object of analysis; second, it demonstrates the flexibility of the proposed framework, as generators can either be represented as satellite agents or be absorbed into the central model, with both configurations seamlessly supported within Annular.

We used Gurobi solver and set `Threads=-1` and `Method=2` so that Gurobi automatically uses the maximum number of threads for maximum parallelization. For the price-response method, we set a common convergence tolerance $\epsilon = 0.1$, applied to both the primal Ψ^k and the dual Ψ^k stopping criteria, and employ the dynamic penalty parameter update strategy of Step 6 of Algorithm 1 to keep the two residuals of comparable magnitude during the iterative process, thereby enabling the use of a single convergence threshold and improving convergence speed. For the market auction method, satellite models are assigned a look-ahead horizon of 24 hours, while the market clearing horizon is set to 1 hour. A 24-hour look-ahead is relatively short for assets with multi-day flexibility, such as hydrogen storage, and may give rise to end-of-horizon artifacts. In such cases, storage can be undervalued near the boundary of the look-ahead window because the satellite optimizer does not fully capture the longer-term value of stored energy beyond the optimization horizon. This limited foresight is expected to contribute to the optimality gap of the market auction approach. Nevertheless, to quantify the magnitude of this effect, we perform a sensitivity analysis in which the look-ahead window is extended, and its impact on the optimality gap is evaluated.

Results and discussion

Accuracy vs. optimality

This section compares the performance of the three coordination approaches in terms of optimality and shows the impact of preserving agents' autonomy (more realistic representation of agent behavior) on the "ideal" optimum computed by co-optimization. We first present the dispatch schedules of agents under each method, and subsequently show the total system operational costs. Finally, we conclude the section by presenting results from a detailed decomposition and sensitivity analysis of the optimality gap between co-optimization and the two model coupling approaches.

Dispatch schedules of agents

The dispatch schedules for all generators, battery storage, and heat prosumers in the experiment with the smallest problem instance are shown in Fig. 5. As expected, the dispatch schedules of agents from co-optimization and price-response are almost identical (see Fig. 7 on the decomposition of the optimality gap). However, the difference in agents' dispatch between market auction and the other two approaches is more pronounced. Take, for instance, the dispatch of electricity generators; a significant difference can be observed in how CHP, gas, and oil-fired generators are used to satisfy peak demand. In the market auction approach, less power is produced by these three generators in earlier periods. However, there is a steep ramp-up in power from these generators thereafter. This difference can be explained through the battery storage dispatch in Fig. 5b. Due to its perfect foresight, full knowledge, and top-down controllability of the MIES, co-optimization can anticipate high prices in the future, and therefore dispatch more production from CHP and gas at an earlier time to charge the battery. The battery is then used to supply peak demand later, thereby preventing the more expensive oil-fired generator from being dispatched at all. The same trend can be observed for the price-response approach, which achieves a very similar dispatch schedule as co-optimization through several iterations.

Conversely, the battery schedule under the market auction approach differs substantially. This difference arises from agents' autonomy and myopic decision-making under imperfect information. Because battery operators retain full autonomy in market auction, they optimize their dispatch based on local objectives, namely revenue maximization. As described in Algorithm 2, the battery operator identifies the lowest forecast price within the horizon and places buy bids at that time in order to discharge at higher prices later. Since the lowest forecast price occurs later in the horizon, this leads to delayed charging and therefore a failure to arbitrage earlier price peaks. As a consequence, early peak demand is instead met by gas, CHP, and oil-fired generators. This behavior reflects the price-taking and imperfect forward-looking limitations of the bidding heuristic in Algorithm 2, which does not internalize market impact or anticipate system-level effects. Therefore, the resulting dispatch differs from the perfect foresight co-optimization benchmark, which can coordinate intertemporal arbitrage optimally by construction. This difference between co-optimization and market auction reflects the characteristic of real-world electricity markets that arise from decentralized decision making and forecast-driven operations under limited foresight.

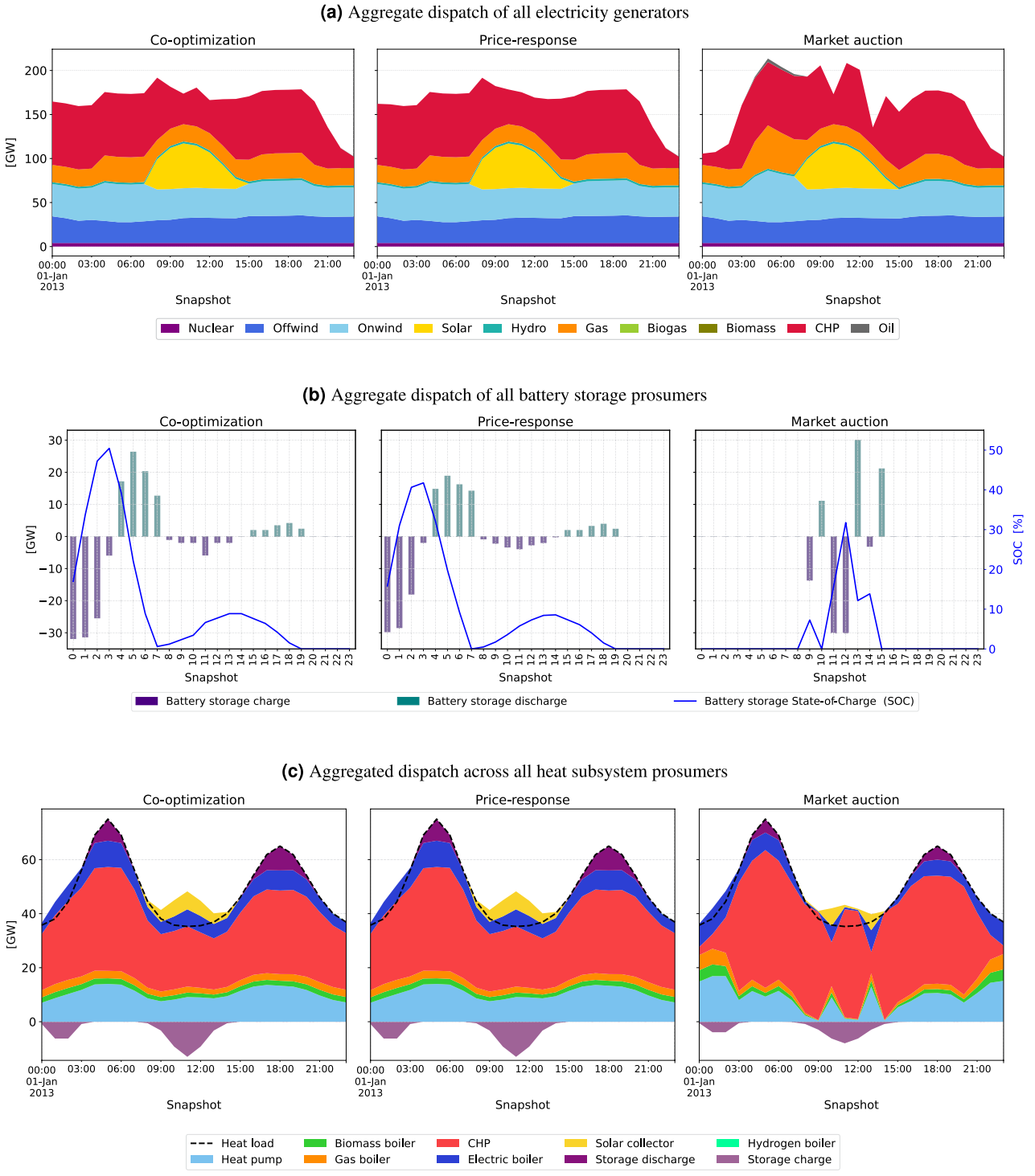


Fig. 5. Dispatch schedules under the three coordination approaches for the smallest problem instance: (**a**) aggregate generator dispatch; (**b**) aggregate battery storage dispatch; (**c**) aggregate heat subsystem dispatch.

Total operational costs and electricity prices

The total operational cost of the MIES and electricity prices are shown in Fig. 6. The total operational cost (see Fig. 6a) under the three approaches is around $20AC314.2$, $20AC314.3$, and $20AC344.8$ million for co-optimization, price-response, and market auction, respectively (see Figure 7 for a detailed analysis of this difference). The higher operational cost under the market auction approach arises directly from the decentralized and forecast-driven dispatch behavior discussed above. In particular, the inability of the battery operator to anticipate and arbitrage the early-morning scarcity period forces the system to rely more heavily on expensive thermal generation. The electricity-price trajectories are also shown in Fig. 6b. Co-optimization and

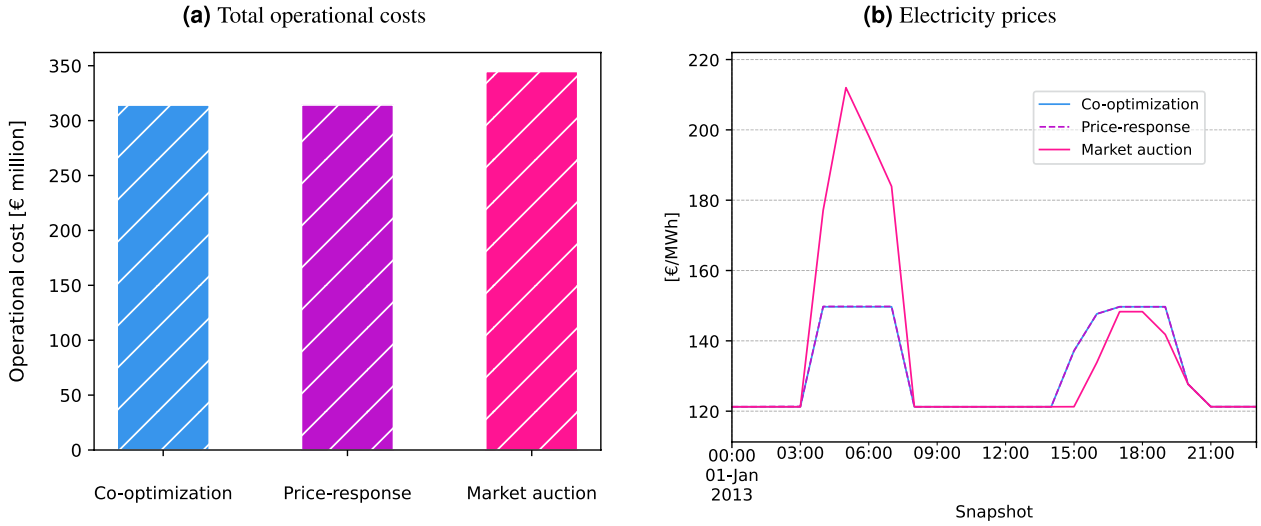


Fig. 6. Total operational cost and electricity price for the smallest problem instance under the three coordination approaches: (**a**) total operational cost of the MIES; (b) hourly market-clearing electricity prices.

price-response produce nearly identical prices, whereas the market auction approach exhibits a similar overall trend, but with substantially sharper price spikes, most notably around 04:00. This spike is directly linked to the battery dispatch behavior discussed above. Because the battery operator delays charging and leaves the battery idle during the early-morning peak (see Fig. 5b), the system lacks low-cost stored energy that could otherwise serve the additional demand. The residual demand must therefore be supplied by the oil-fired generator, which becomes the marginal unit during that hour. Since market-clearing prices are set by the marginal generator, the dispatch of the oil-fired unit directly causes the electricity price spike to approximately 210 $20AC/MWh$ in Fig. 6b.

It is worth noting that the costs and prices obtained from co-optimization should only be interpreted as a theoretical reference benchmark and not as a strict performance metric, since they result from an "ideal" setting characterized by perfect foresight, perfect competition, and top-down centralized dispatch; whereas actual market outcomes are almost never optimal in reality. Accordingly, the objective of the market auction approach is not to minimize prices directly (see Fig. 13 on the nuance in model selection), but rather to provide a modeling environment capable of fully representing the behavior and interactions of flexibility providers. This, in turn, provides a testbed for evaluating how alternative policy regulations and new market design mechanisms influence the behavior of flexibility providers and their resulting impacts on system-wide social welfare.

In sum, our findings show that while both co-optimization and price-response can find the theoretical system-wide optimum, they are, however, less accurate (with respect to modeling flexibility under imperfect competition) since both methods disregard features such as the myopic, bounded-rationality bidding behavior of self-interested flexibility providers that are inherent in real-world decentralized market settings. On the other hand, while market auction does not achieve the same "ideal" optimum as the other two approaches, it can fully capture these key features of flexibility that must be studied and understood— in order to avoid myopic operational and investment decisions for market parties, and for regulators to investigate policy instruments that can lead to an efficient integration and utilization of flexibility resources in MIES.

Decomposition and sensitivity analysis of the optimality gap

The total operational costs of the MIES reported in Fig. 6 correspond to the smallest problem instance (30 agents, 24 h simulation horizon) under the default settings of the market auction approach: a 24 h look-ahead window and a step-function bid resolution of 5 price—quantity blocks. Under these settings, the market auction approach exhibits an approximately 9.7% higher operational cost than co-optimization, which can be interpreted as the "theoretical" optimality gap between the two approaches. However, two questions remain unanswered: (i) does this gap persist across different problem instances? and (ii) which modeling choices drive the gap, and to what extent?

To address these questions, we perform a decomposition and sensitivity analysis of the objective-cost deviation across all 44 problem instances (4 simulation horizons $\times$ 11 agent counts). Figure 7 reports the optimality gap (percentage deviation in total operational cost relative to co-optimization) under three experimental conditions: (i) the baseline configuration (24 h look-ahead, 5 bid blocks); (ii) a higher bid resolution variant (10 bid blocks, look-ahead unchanged); and (iii) a longer look-ahead variant (168 h look-ahead, bid resolution unchanged). These two sensitivity dimensions apply only to the market auction approach and therefore do not affect the price-response results; consequently, the left-column panels in Fig. 7 are identical across all three rows, and only serve to show the optimality gap across different problem instances in the price-response approach.

Two observations stand out. First, the price-response approach (left column of Fig. 7) exhibits an optimality gap below 0.1% across all 44 problem instances. The results are identical across the three rows because the price-

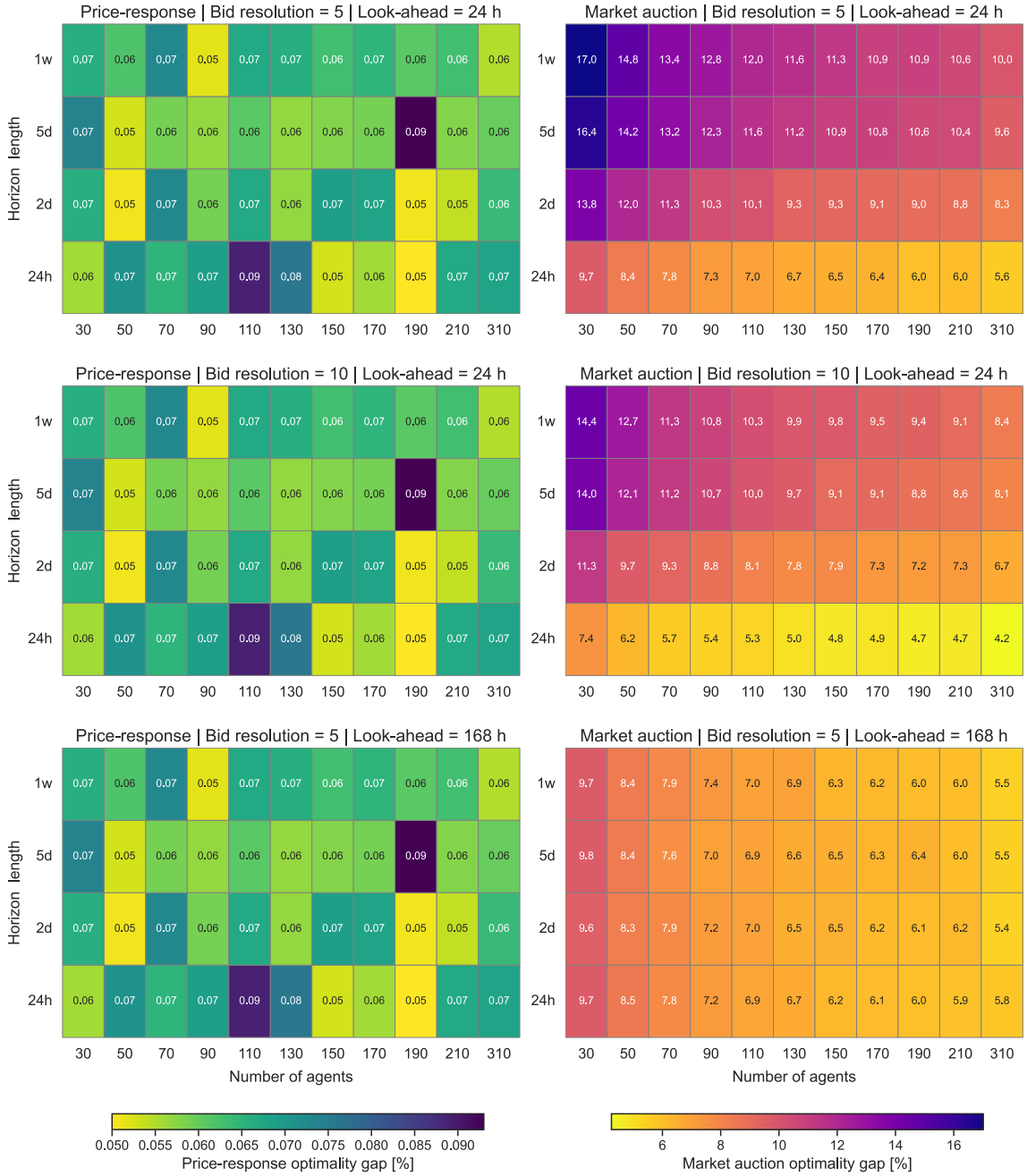


Fig. 7. Decomposition and sensitivity analysis of the optimality gap (% deviation in total operational cost from co-optimization) for the price-response (left column) and market auction (right column) approaches, across all 44 convex problem instances (4 simulation horizons × 11 agent counts) under three experimental conditions (rows): a baseline (look-ahead = 24 h, bid resolution = 5 blocks); a higher-resolution bid variant (10 blocks, look-ahead unchanged); and a higher-look-ahead variant (168 h, bid resolution unchanged). Cells are annotated with the gap value. The two variants are knobs of the market auction approach and do not apply to the price-response approach: the price-response panels (left column) are therefore identical across all three rows and simply report the price-response optimality gap across instances; they are repeated in rows 2 and 3 only to allow a per-row visual comparison against the market auction panel on the right.

response approach does not rely on rolling-horizon look-ahead or bid functions, and is therefore unaffected by either sensitivity dimension. This empirically confirms the theoretical expectation that price-response recovers the centralized co-optimization solution within the predefined convergence tolerance for convex problems. Second, the market auction approach (right column of Fig. 7) exhibits a substantially larger and more variable gap, which can be decomposed into the following three main effects.

(i) Foresight effect. Extending the look-ahead window from 24 h to 168 h reduces the gap in the worst-case instance (30 agents, 1-week simulation horizon) from approximately 17% to 9.7%, corresponding to a reduction

of about 7 percentage points. This quantifies the effect of imperfect foresight. With a 24 h look-ahead applied to a 168 h simulation, satellites optimize only over a limited future window at each commitment step. When the look-ahead is extended to cover the full simulation horizon, satellites effectively operate under perfect foresight and this component of the gap largely disappears.

(ii) Bid resolution effect. Increasing the bid resolution from 5 to 10 price—quantity blocks reduces the gap in the baseline instance (30 agents, 24 h horizon) from approximately 9.7% to 7.4%, a reduction of about 2.3 percentage points that is relatively consistent across instances. This reflects the approximation error introduced by representing the satellite's underlying cost curve with a coarse step-function bid. A higher bid resolution provides a more granular representation of agents' underlying marginal cost and willingness-to-pay curve. This allows the market to reflect finer variations in marginal valuation that would otherwise be aggregated within coarser bid blocks, leading to a more efficient market-clearing outcome.

(iii) Agent averaging effect. Across all market auction panels, the optimality gap decreases as the number of agents increases. Under the baseline configuration, the gap falls from approximately 9.7% at 30 agents to 5.6% at 310 agents, while under the higher bid resolution configuration it decreases from 7.4 to 4.2%. This reflects an averaging effect: as the number of participating agents grows, the individual imperfections and myopia of local bidding heuristics become partially averaged out through aggregate market clearing, resulting in more efficient system-level outcomes.

Together, these three effects explain most of the variation in the market auction optimality gap. The worst-case configuration (30 agents, 1-week horizon, 24 h look-ahead, 5 bid blocks) reaches approximately 17%, whereas the best-case configuration (310 agents, 24 h horizon, 24 h look-ahead, 10 bid blocks) falls to approximately 4.2%. The remaining 4–5% gap, therefore, represents the residual efficiency cost inherent to decentralization, even under favorable modeling assumptions. Notably, this residual range is comparable to the 1–5% efficiency-loss range commonly reported for well-functioning wholesale electricity markets [60,61], suggesting that the efficiency cost of fully decentralized coordination is of similar magnitude to inefficiencies already observed in decentralized electricity markets.

Computational efficiency and scalability

In this section, we compare the computational efficiency (reflected by absolute runtime) and scalability (reflected by normalized runtime) of the three coordination approaches under convex and non-convex problem instances. We also discuss the runtime composition (the time spent on computation, communication, and waiting) of price-response and market auction. Finally, we conclude with an analysis of peak memory usage (reflected by Maximum Resident Set Size, RSS) across the three approaches.

Runtime and scalability

The computational efficiency and scalability results for the convex problem instances are shown in Fig. 8. We start with the computational efficiency reflected by the cumulative distribution of absolute runtime across all convex problem instances (Fig. 8a). Each problem instance—defined by a unique combination of number of agents and horizon length—is run as a separate job. Therefore, Fig. 8aconveys two types of information: (i) the total time required to complete all problem instances under different experimental settings for each approach; and (ii) the speed of computations, expressed as the proportion of solved problem instances as a function of wall-clock time.

The following observations can be made from Fig. 8a. First, the market auction approach is the fastest overall, as indicated by its lowest cumulative absolute runtime across all experiments. Price-response outperforms co-optimization only when more CPU cores are available. The practical implication is that the relative ranking of the three approaches is hardware-dependent: on a 16-core machine (the leftmost panel of Fig. 8a), the price-response approach is in fact the slowest of the three across most of the problem instance distribution, and only becomes competitive with co-optimization once more CPU cores are available to parallelize the per-agent subproblems. From the steep initial rise of its cumulative distribution curve, co-optimization is fastest for small problem instances, as it avoids coordination overhead associated with iterative updates and communication overhead among models. However, the runtime of co-optimization increases sharply as problem size grows, leading to significantly slower performance for larger problem instances. Therefore, while co-optimization is efficient for small-scale problems, it becomes computationally inefficient as the problem size increases.

Second, increasing the number of CPU cores results in significant speedup (lower runtime) for both the price-response and market auction approaches, as reflected by the leftward shift of their curves across subplots. This effect is most pronounced for the market auction approach; market auction improves cumulative batch throughput by approximately 30% on the 16-core configuration, and by up to 70% on the 128-core HPC node relative to co-optimization. These gains are batch-throughput gains across all problem instances rather than per-instance speed-ups; the per-instance speed-up at the 128-core level is approximately a factor of 2–3 for larger problem sizes. For co-optimization, an initial increase in the number of CPU cores from 16 to 48 leads to a moderate speedup. However, further increasing the number of CPU cores from 48 to 128 actually results in longer co-optimization runtime due to diminishing returns at higher parallel thread counts. That is, additional threads duplicate the working memory, and the per-thread synchronization overhead eventually offsets parallelization gains. This highlights the extent to which different coordination architectures can efficiently exploit parallel computing resources: the distributed and decentralized architectures of model coupling scale naturally with additional cores because they decompose the problem into many smaller subproblems that can be solved concurrently, whereas the monolithic co-optimization formulation exhibits limited parallel scalability due to the centralized structure of the problem definition.

To further assess the scaling trend of each approach, we normalize the runtime of each problem instance by the runtime of the smallest instance for each approach. Normalization also eliminates any bias from implementation-

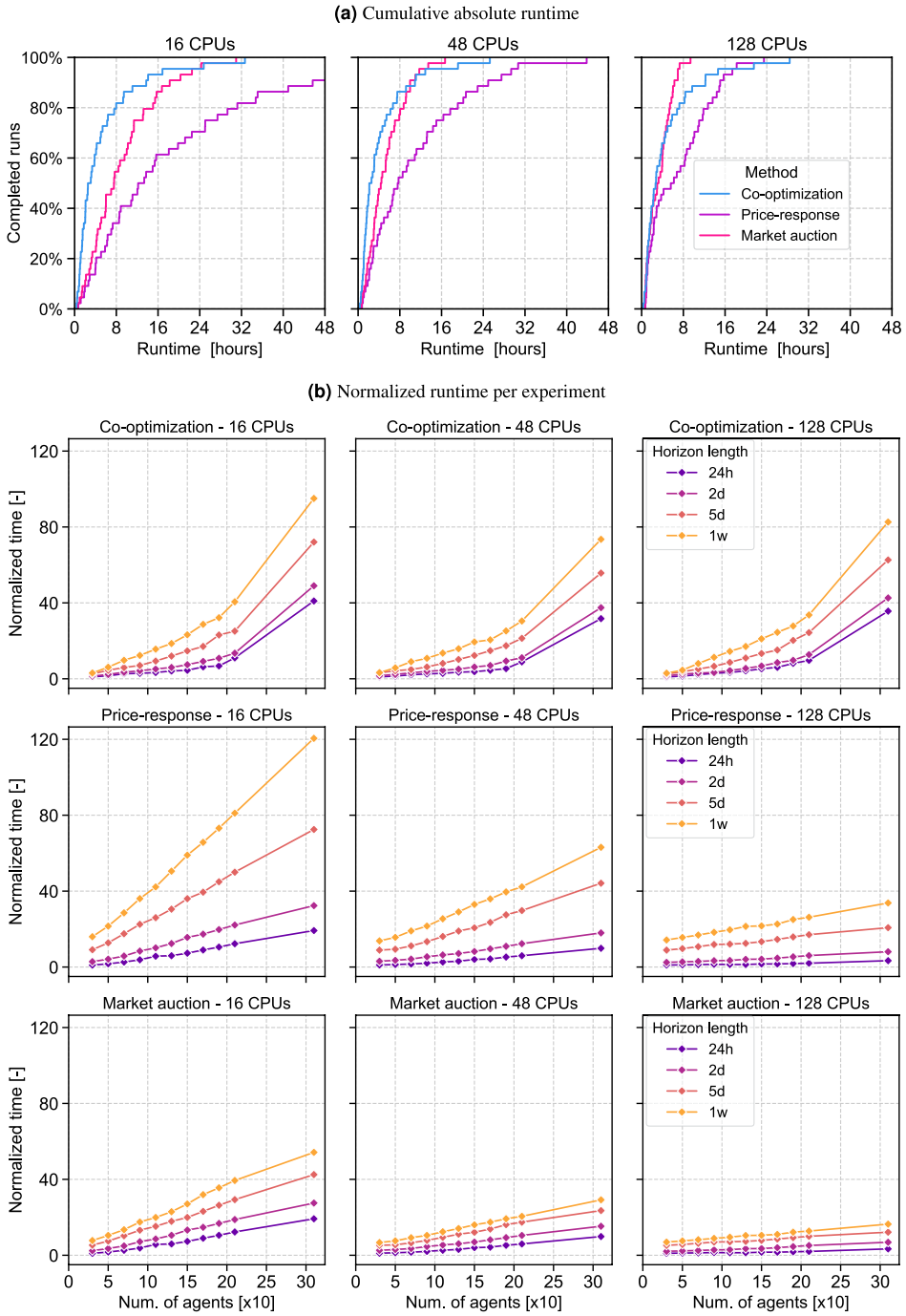


Fig. 8. Scaling performance of the three coordination approaches across the 44 convex (LP) problem instances (4 simulation horizons × 11 agent counts), run at three CPU configurations (16, 48, 128 cores): (**a**) empirical cumulative distribution of wall-clock runtime; (**b**) per-method minimum-time-normalized runtime.

specific optimizations, since each method is compared against its own baseline. The result of the scaling trends for convex problem instances is shown in Fig. 8b. We see that under fixed computational resources, the solution time for co-optimization increases rapidly with problem size at an approximately polynomial rate ($O(n^2)$ in the agent count n), consistent with the typical behavior of barrier and simplex solver methods on this class of problems, whereas both price-response and market auction scale almost linearly. Among these, the market auction approach scales substantially faster as reflected by its lowest slope. Consistent with the cumulative distribution of absolute runtime (see Fig. 8a) discussed earlier, we again see that as computational resources increase, co-optimization shows a moderate improvement in speedup up to 48-cores, and then slows down as the number of CPU cores increases further. On the other hand, price-response and market auction benefit significantly from parallelism, as indicated by the reduced slopes of their curves across subplots.

The difference in scalability can be explained by differences in the coordination architecture, which determines how efficiently each modeling approach can exploit parallelism (as discussed above), and the coordination mechanism, which determines the number of computational steps required per solution. Co-optimization exhibits the weakest scaling performance since it relies on a monolithic problem structure that offers limited opportunities for decomposition and parallel execution. The market auction approach is substantially faster than the price-response approach because it is non-iterative—that is, the interaction between the central and satellite models is executed once per market-clearing round. In contrast, the iterative nature of price-response introduces a computational bottleneck since agents must repeatedly solve their local flexibility scheduling problems until convergence (see Figs. 9aand 9b). This implies that the total solution time of price-response increases proportionally to the number of iterations required for convergence. Moreover, while price-response only decomposes the problem across space (agents), market auction decomposes the problem across both space (agents) and time, since it partitions the horizon into smaller rolling windows and solves them sequentially in a rolling manner.

We additionally examine whether the iteration count of the price-response scales with problem size. Figure 9creports the iteration count to convergence across all 44 convex problem instances. The mean iteration count is about 1000, with per-instance counts ranging from approximately 700 to 1300 and showing no monotonic trend with respect to either the agent count or horizon length. Therefore, the convergence behavior of the price response is dominated by its consensus dynamics rather than the dimensionality of the underlying problem, implying that larger instances do not necessarily require more iterations but instead require more computational effort per iteration.

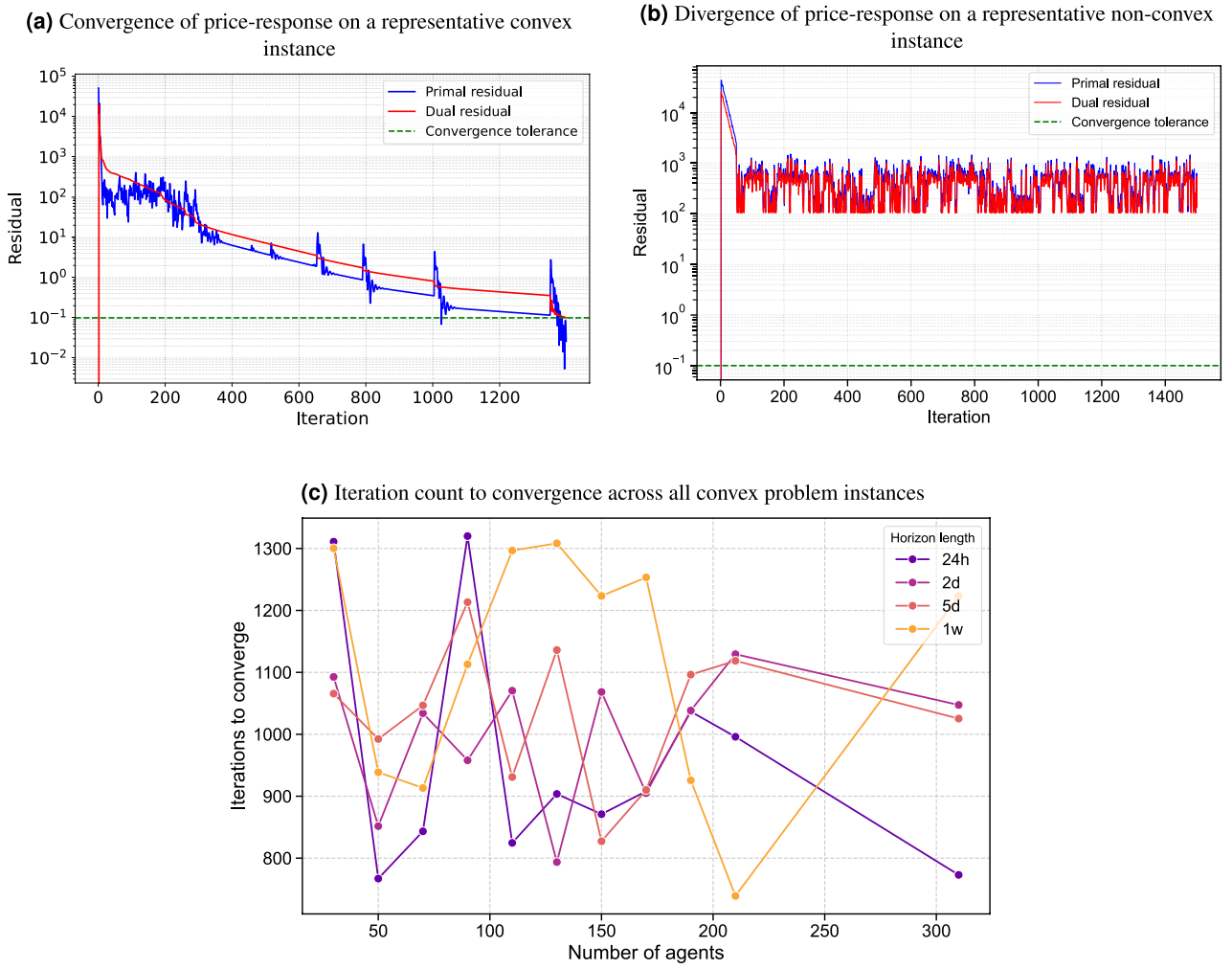


Fig. 9. Analysis of the convergence behavior of price-response. (**a**) Primal and dual residuals decay to the convergence tolerance for a representative convex instance. (**b**) On a representative non-convex instance, both residuals enter a sustained flip-flop state after an initial fast decay and never cross the tolerance. (**c**) Iteration count to convergence across all 44 convex problem instances (4 horizons × 11 agent counts); the mean is 1000, with no monotonic scaling trend in either dimension.

Behavior under non-convexity

The results discussed so far are for the convex (LP) setting. To analyze the scaling behavior of the three coordination approaches under non-convexities (MILP), we repeat all experiments with the storage complementarity binaries introduced in the Experiment Design section. Figure 10 shows the empirical findings on runtime and convergence behavior.

The cumulative distribution of absolute runtime in Fig. 10a reveals that both co-optimization and the market auction approach remain robust under non-convexities, completing all problem instances within the 48-hour wall-clock budget on all CPU configurations. Their runtime are consistently higher than in the convex case, indicating that the introduction of storage complementarity increases computational effort but does not fundamentally alter the relative efficiency ordering established earlier above for the convex cases. In particular, the market auction approach continues to reach full completion earlier than co-optimization across all CPU settings. In contrast, the price-response approach exhibits a substantial loss of robustness under non-convexity, completing only around 27% of all instances within the same wall-clock budget. This fraction increases only marginally with additional CPU resources (from 22.7% at 16 cores to 29.5% at 128 cores), reflecting an inherent convergence limitation rather than a purely computational one.

The scaling behavior shown in Fig. 10b further confirms that the relative scaling structure of co-optimization and market auction is essentially unchanged compared to the convex cases (Fig. 8b). This invariance arises because the per-method minimum-time normalization removes the uniform slowdown induced by the non-convexities. Consequently, both approaches preserve their scaling trend, with the market auction approach still being the fastest (the smallest slope). The price-response scaling curves, in contrast, are sparse because only the subset of instances that converge within the wall-clock budget contribute runtime observations; diverged instances are excluded since they do not yield any meaningful wall-clock runtime.

The limited completion rate of price-response can be traced to its convergence dynamics, as illustrated in Fig. 9b. After an initial phase of rapid decay in primal and dual residuals, similar to the convex case, the algorithm enters a persistent oscillatory state in which primal and dual residuals fluctuate without approaching the convergence threshold. This behavior is driven by discrete state transitions in the binary storage variables, which induce abrupt changes in individual agent responses as the price signal evolves. While the dynamic-ρ update rule mitigates imbalance when residuals become large, it cannot suppress these discontinuous flips, leading to sustained non-convergent cycles.

Overall, the non-convex results in Fig. 10 confirm the trends observed in the convex setting while highlighting a key distinction in robustness: co-optimization and market auction remain computationally stable under non-convexities, whereas price-response loses a substantial fraction of solvable instances due to convergence failure. This reinforces the interpretation that market auction is the only coordination architecture among the three that combines scalability with robustness to non-convex operational constraints, while co-optimization remains reliable but computationally more demanding, and price-response is sensitive to discrete state dynamics in the underlying optimization problems.

Decomposition of runtime

To give a better indication of time spent in model communication (information exchange) and actual computation, we decompose the wall-clock runtime of the two model coupling approaches into computation, communication, and waiting components. Figure 11 reports the per-instance wall-clock breakdown for both the price-response and market auction approaches on the smallest problem instance. For both approaches, the communication share is very small (approximately 1.8% for price-response and 1.0% for market auction); runtime is overwhelmingly dominated by the local computations of the slowest agents and the waiting time of the fastest agents, rather than by inter-model data exchange (communication). In both model coupling approaches, the slowest agent (P14) spends more than 95% of its runtime on local computation, while the fastest agents (the coordinator agent in price-response and G16 in market auction) spend less than 5% on computation, remaining idle for most of the simulation while waiting for slower agents to complete.

What makes a particular submodel slow in our experiments is primarily the size of its asset portfolio and, most importantly, whether it owns storage and/or flexible demand (modeled as a store, as described in the "Methodology" section). Storage assets and flexible demand introduce additional variables and constraints with strong intertemporal coupling across the optimization horizon. This translates directly into longer solve times and explains the heavy-tailed compute distribution visible in Fig. 11: the slowest agents in both Figs. 11aand 11bare prosumers that own both battery and thermal storage, whereas the fastest agents neither own storage nor solve computationally intensive scheduling problems. In the price-response approach, the coordinator only performs imbalance aggregation and price updates, while in the market auction approach, G16 submits a single bid block equal to its available capacity at marginal cost.

Therefore, the overall runtime of both model coupling approaches is ultimately constrained by the slowest submodel. This implies that further improvements in computational efficiency can effectively be achieved by targeting the computational bottlenecks of the slowest submodel, for example, through model reduction, temporal aggregation, warm-starting, decomposition of storage scheduling, or surrogate representations of computationally intensive flexibility assets.

Memory usage

Beyond runtime, scalability also depends on memory usage, since memory exhaustion is a common failure mode for large-scale monolithic energy system optimization models. Figure 12 reports the peak memory usage (Maximum Resident Set Size, RSS) of the three coordination approaches across all convex problem instances and CPU configurations. We do not show a separate non-convex memory figure because the introduction of non-convexities (through storage-complementarity binaries) does not materially change the qualitative memory

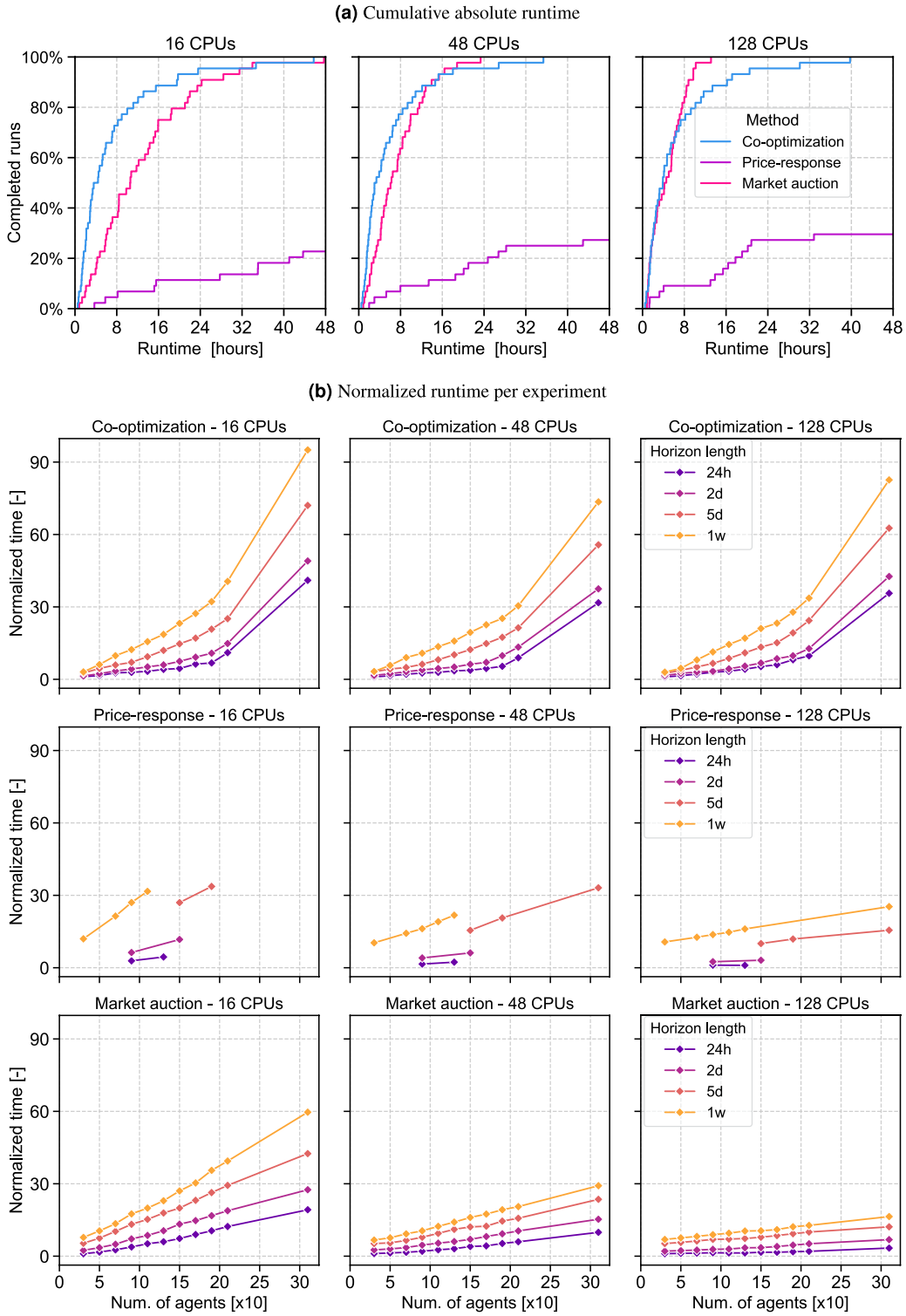


Fig. 10. Scaling performance under the storage charge–discharge complementarity non-convexity, across the 44 non-convex (MILP) problem instances and three CPU configurations: (**a**) empirical cumulative distribution of wall-clock runtime; (**b**) per-method minimum-time-normalized runtime. Price-response converges within the wall-clock limit on only 27% of instances on average; only converged runs contribute points in (**b**), and diverged instances are omitted.

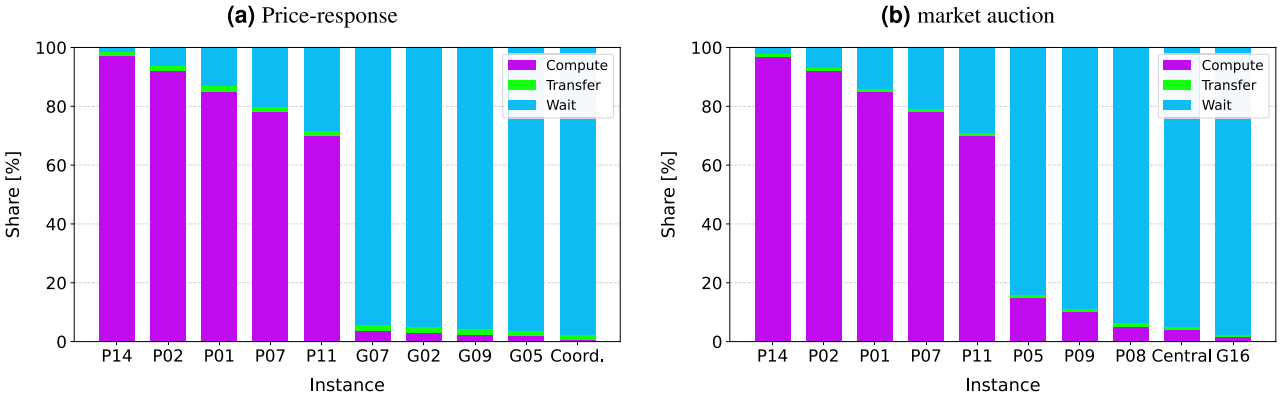


Fig. 11. Per-instance share of wall-clock time spent in compute, transfer, and wait for (**a**) price-response and (**b**) market auction averaged over all convex runs. Each plot shows the five highest-compute and five lowest-compute instances. *Coord.* coordinator agent in the price-response approach; *Central* central market-clearing in the market auction approach.

scaling behavior observed in the convex case: co-optimization remains the most memory-intensive approach, while the market auction approach exhibits the slowest memory growth.

Two trends stand out from Fig. 12. First, peak memory usage increases with problem size for all three approaches, as expected. Second, and more importantly, peak memory also increases with the number of CPU cores allocated to the solver, but the rate of increase differs substantially across the three coordination architectures. This growth is steepest for co-optimization, more moderate for the price-response approach, and smallest for the market auction approach. The mechanism behind this pattern is tied to how the solver (Gurobi's Barrier solver method) parallelizes computations. Under the Barrier method (used in our experiments; see the "Experiment design" section), each additional thread allocates its own working memory for the sparse Cholesky factorization of the optimization problem[62]. The relevant problem, however, differs substantially across the three coordination approaches. In co-optimization, all threads operate on a single large monolithic factorization, so each added thread compounds the memory requirement of the same large problem. In the price-response approach, the problem is decomposed across agents, so each thread works on a much smaller per-agent subproblem with correspondingly smaller working memory. The market auction approach decomposes the problem even further by limiting each satellite optimization to a rolling time window, thereby producing the smallest per-thread working set among the three approaches.

In other words, the deeper the decomposition of the optimization problem (across agents and across time), the smaller the per-thread memory footprint and the slower the growth in memory consumption as additional CPU cores are introduced. These results reinforce the runtime analysis above: the model coupling approaches—and especially the market auction approach—not only scale better computationally, but are also substantially less susceptible to memory exhaustion at large problem sizes.

Nuance in model selection

So far, we have examined the three coordination approaches across several dimensions: optimality, autonomy, privacy, scalability in both runtime and memory, and behavior under non-convexities. Taken together, these findings do not identify a single "best" approach; rather, they reveal important trade-offs that make the choice of coordination architecture dependent on the modeler's objectives and constraints. We now synthesize the empirical evidence presented thus far into practical recommendations for energy system modelers. To do so, we develop a decision tree (Fig. 13) that guides the selection of a coordination approach through five sequential questions, ordered by how decisive the answer tends to be.

At a high level, the recommendation is as follows. If minimum information disclosure is a hard requirement (Q1)—for example, when modeling residential prosumers whose techno-economic parameters and internal states cannot be shared with a central planner—co-optimization is ruled out, and the final choice between price-response and market auction is deferred to the remaining questions below. If full local autonomy in agents' decision making is required (Q2)—for example, when studying strategic bidding, market power, or market mechanism design—market auction is the only architecture that satisfies this constraint; however, this comes at the cost of a measurable optimality gap (approximately 4–5% in the best case, and up to 17% in the worst case, see Fig. 7). When non-convexities are present in the problem definition (Q3)—for example, storage charge–discharge complementarity— the price-response approach should be avoided since only approximately 27% of MILP instances converge within our wall-clock budget; the choice then collapses to co-optimization or market auction depending on problem size and available computing resources (Q4). When none of the above constraints applies, and the problem is convex, the choice is driven by the ratio between problem size and available computing resources: for small problems, or for larger problems on modest hardware, co-optimization remains the most efficient option since neither model coupling approach can outpace it without parallel cores (Q4); for larger problems on HPC infrastructure (> 16 cores with ample memory), the choice between price-response and market auction is governed by the optimality tolerance the modeler is willing to accept (Q5)—

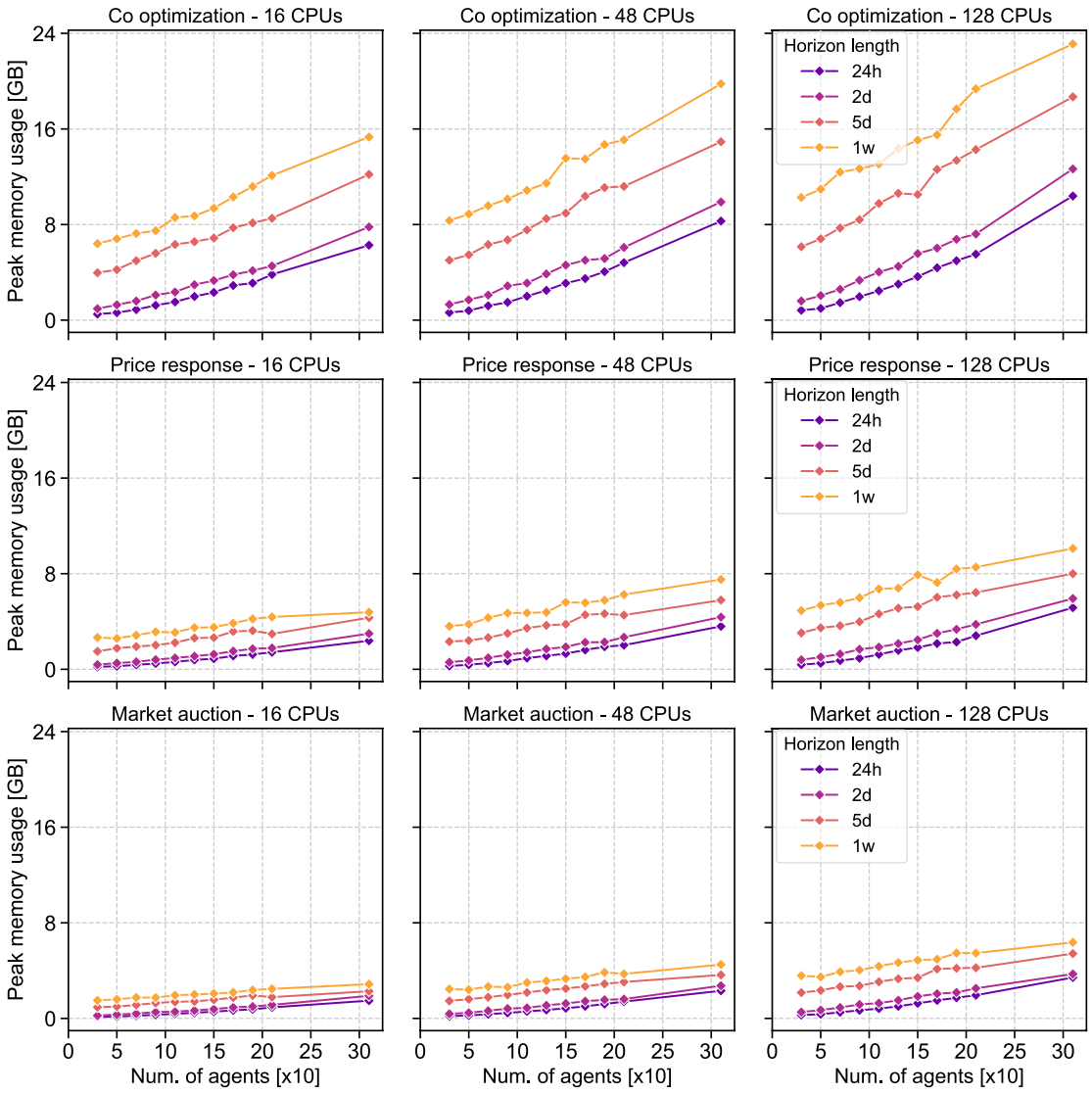


Fig. 12. Peak memory usage (maximum resident set size, RSS) of the three coordination approaches across all 44 problem instances and three CPU configurations. Memory grows with both problem size and the number of CPU cores allocated to the solver; the growth rate with cores is steepest for co-optimization, more moderate for the price-response approach, and slowest for the market auction approach.

price-response when a tight ($\leq$1% gap) match with the centralized optimum is required, market auction when a $\geq$5% gap is acceptable in exchange for the best runtime and memory scaling of the three approaches.

Transferability of findings to spatially-resolved systems

The benchmark presented above adopts a copper-plate representation of the electricity system, in line with both European zonal market design and our objective of isolating the coordination layer from confounding network effects as introduced earlier in the "Methodology" section. Similarly, the recommendations from Fig. 13 are also based on the same copper-plate electricity system assumption. Here, we discuss the potential changes when this assumption is relaxed, the boundaries within which our quantitative claims should be interpreted, and the implications for the recommendations presented above.

Network constraints alter the coordination problem along three dimensions. First, the market-clearing price ceases to be a one-dimensional (time-indexed) vector and becomes a two-dimensional (time- and location-indexed) vector of locational marginal prices, so the information-exchange protocol of every coordination approach must be extended accordingly. For co-optimization, this is mechanical: the system-wide power balance (Eq. 2) is replaced by a set of nodal balances coupled by line-flow constraints, and the cost-optimal dispatch is recovered from a larger but structurally similar problem. For the price-response approach, the consensus variable becomes a nodal time-indexed imbalance vector; the algorithm extends naturally to this setting, but at the cost of a higher-dimensional consensus problem and, in general, an expected slower convergence. For the market

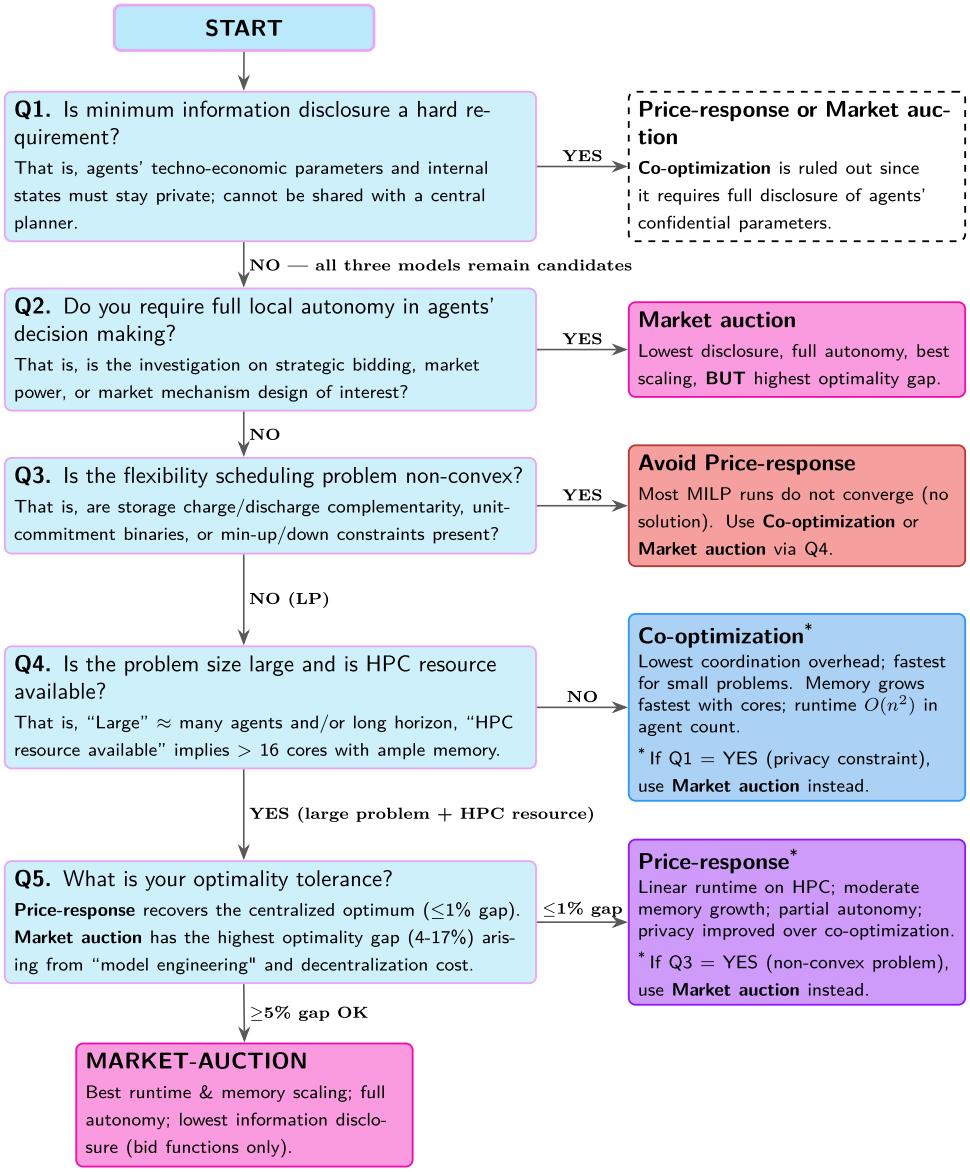


Fig. 13. Decision tree for selecting a method for modeling flexibility scheduling in MIES. Five sequential questions route the modeler to one of three recommended flexibility modeling approaches: co-optimization, price-response, or market auction. Q1 (privacy) rules out co-optimization when minimum information disclosure is required and defers the final choice to the remaining questions; Q2 (autonomy) routes directly to market auction when full agent autonomy is required in the model; Q3 (non-convexity) flags price-response as unreliable for MILP problems and routes back through Q4; Q4 (problem size and HPC availability) selects co-optimization for small problems or modest hardware and otherwise continues to Q5; and Q5 (optimality tolerance) resolves the final choice between price-response and market auction. Asterisked leaves carry footnotes that override the default recommendation when an upstream constraint applies.

auction approach, the central market-clearing must be replaced by a network-constrained clearing analogous to standard nodal locational marginal pricing markets, while the satellite bid functions can, in principle, remain unchanged provided agents' bids are interpreted at their connection nodes.

Second, the feasible dispatch region of each individual agent becomes coupled to the network state reflected by the level of utilization (or congestion). In small problem instances with few large agents, this can introduce strategic incentives that the price-taking bidding heuristic in Algorithm 2 does not capture. The qualitative finding that the market auction approach trades a measurable optimality gap for full local autonomy is, however, expected to persist; if anything, the gap is likely to grow under network constraints because locational price signals create additional opportunities for myopic dispatch decisions to diverge from the system optimum.

Third, our scalability claims should be interpreted as upper bounds on the speedup achievable in the network-constrained case. The decomposition advantage of the market auction approach—which partitions the problem across both space (agents) and time (rolling market windows)—relies on the fact that, within a clearing

window, satellite problems are independent. Adding network constraints reintroduces a coupling between satellites at the central clearing step, but does not destroy the decomposition: the satellite problems themselves remain independent and parallelizable. The price-response approach is more directly affected, since its iterative consensus loop must now reach agreement on a time and location-indexed vector. Consequently, we expect the relative ranking of the three approaches by absolute runtime to be preserved on hardware with sufficient parallelism, but the magnitude of the speedup reported here (30% on a 16-core machine and up to 70% on a 128-core machine) is specific to the copper-plate setting and should not be extrapolated quantitatively to spatially-resolved studies.

In sum, the conceptual conclusions of this work—that model coupling provides a viable paradigm for preserving the autonomy and privacy of flexibility providers, that the price-response approach recovers the centralized optimum at the cost of iterative overhead, and that the market auction approach offers the most favorable scaling behavior at a measurable optimality cost—are expected to carry over to network-constrained settings. The specific numerical claims are bounded to the LP/MILP, copper-plate setting examined here, and we leave a quantitative assessment under network constraints to future work.

Conclusion

The increasing penetration of distributed energy resources and the resulting interdependencies across energy carriers are transforming energy systems into highly complex, multi-carrier integrated energy systems (MIES). While the large-scale integration of distributed energy resources offers new sources of flexibility to cope with the variability of renewable energy, such as wind and solar, and to manage grid congestion, coordinating flexibility across many heterogeneous, autonomous prosumers remains a fundamental methodological challenge. Conventional monolithic co-optimization, while theoretically optimal, fails to adequately capture the autonomy and preserve the privacy of flexibility providers—who are increasingly participating in active demand-response—and faces significant scalability limitations due to its centralized coordination architecture.

In this work, we investigated runtime model coupling as an alternative paradigm for modeling the coordination of flexibility scheduling in MIES. We argued that model coupling can overcome the limitations of monolithic approaches by enhancing the autonomy and privacy of flexibility providers—since coupled models locally optimize their decisions and only share interface variables—and also promises better scalability since coupled models can be simulated in parallel. We introduced two model coupling approaches: price-response and market auction, which can be used to couple many energy models at runtime to effectively capture the dynamic, mutually dependent interactions among flexibility providers. These two approaches mainly differ in their coordination architecture, coordination mechanism, information exchange protocol, and the extent to which autonomy is retained in agents' local decision-making. We performed a systematic quantitative benchmark of these two approaches with monolithic co-optimization under varying problem sizes, complexities, and computing infrastructures.

The results reveal significant trade-offs between optimality, autonomy, and scalability. Co-optimization can compute the theoretical system-wide optimum and performs well for small problem instances, but scales poorly as the number of agents and the temporal dimension increase. The price-response model coupling approach improves scalability while preserving a notion of coordinated optimality. However, its iterative nature introduces computational overhead that limits performance gains, particularly for large-scale problems with long time series, and faces convergence issues in the presence of non-convexities. In contrast, the market-auction approach achieves the most favorable scalability (even under non-convexities) through the decomposition of the problem in space (agents) and time, and its non-iterative coordination mechanism; however, the speedup comes at the cost of "theoretical" optimality. Nevertheless, modeling coupling offers an approach to balance realism (preservation of agents' autonomy and privacy) with computational tractability, improving cumulative batch throughput by approximately 30% on a 16-core machine and up to 70% on a 128-core HPC node, relative to co-optimization. We emphasize that the reported quantitative gains are specific to the copper-plate LP/MILP benchmark considered in this study, and therefore, more work is needed to assess the extent to which these findings generalize to network-constrained formulations with nonlinear AC power flow models.

Nevertheless, the findings we uncover here advance the discussion on modeling future energy systems dominated by distributed energy resources by providing both conceptual and practical guidance, supported by empirical evidence, on how to coordinate flexibility scheduling in increasingly decentralized and integrated energy systems. By conducting a systematic quantitative benchmark of different coordination approaches, we provide empirical insights and recommendations to help energy system modelers select the coordination model that matches their modeling priorities and constraints, model scale and complexity, and available computing resources. In particular, our results suggest that co-optimization remains the preferred choice for small-scale problems and for settings in which preserving agents' autonomy and privacy is not a hard constraint; price-response is most suitable when partial autonomy in agents' decision-making, minimal information disclosure, and textbook optimality outcomes are desired, provided the problem remains convex; and market auction is preferable when full autonomy in agents' decision-making, minimal information disclosure, scalability, robustness to non-convexities, or the study of market behavior and policy design are primary modeling objectives, and near-optimality is acceptable. Beyond introducing runtime model coupling as a new paradigm for modeling flexibility in MIES, this work also helps avoid time-intensive trial-and-error exploration by offering clear, evidence-based guidance on the strengths, limitations, and trade-offs of different coordination approaches across both theoretical and practical performance metrics. Finally, to support further research and application, the proposed approaches have been formalized and implemented as open-source software tools, enabling the energy modeling community to reproduce, extend, and adapt these approaches in their own direction of interest.

Data availability

The code to reproduce the experiment results and visualizations in this study is available on GitLab: https://gitlab.tudelft.nl/demoses/demoses-scaling-benchmarks, and is archived on Zenodo: https://zenodo.org/records/15496935

Received: 26 March 2026; Accepted: 8 June 2026
Published online: 13 June 2026

References

1. Mancarella, P. MES (multi-energy systems): An overview of concepts and evaluation models. *Energy* **65**, 1–17. https://doi.org/10.1016/j.energy.2013.10.041 (2014).
2. Martínez Ceseña, E. A., Loukarakis, E., Good, N. & Mancarella, P. Integrated electricity– heat–gas systems: Techno–economic modeling, optimization, and application to multienergy districts. *Proc. IEEE* **108**, 1392–1410. https://doi.org/10.1109/JPROC.2020.2989382 (2020).
3. Garg, A., Niazi, K. R., Tiwari, S., Sharma, S. & Rawat, T. Optimal energy management of multi-carrier energy system considering uncertainty in renewable generation. *Sci. Rep.* **15**, 25936. https://doi.org/10.1038/s41598-025-10404-4 (2025).
4. O'Connell, N., Pinson, P., Madsen, H. & O'Malley, M. Benefits and challenges of electrical demand response: A critical review. *Renew. Sustain. Energy Rev.* **39**, 686–699. https://doi.org/10.1016/j.rser.2014.07.098 (2014).
5. Poncelet, K., Delarue, E. & D'haeseleer, W. Unit commitment constraints in long-term planning models: Relevance, pitfalls and the role of assumptions on flexibility. *Appl. Energy* **258**, 113843. https://doi.org/10.1016/j.apenergy.2019.113843 (2020).
6. Li, J. et al. An optimization method for integrated demand response strategies for electricity and heat considering the uncertainty of user-side loads. *Sci. Rep.* **16**, 582. https://doi.org/10.1038/s41598-025-30090-6 (2025).
7. Mohsenian-Rad, A.-H., Wong, V. W. S., Jatskevich, J., Schober, R. & Leon-Garcia, A. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Trans. Smart Grid* **1**, 320–331. https://doi.org/10.1109/TSG.2010.2089069 (2010).
8. Ghatikar, G., Mashayekh, S., Stadler, M., Yin, R. & Liu, Z. Distributed energy systems integration and demand optimization for autonomous operations and electric grid transactions. *Appl. Energy* **167**, 432–448. https://doi.org/10.1016/j.apenergy.2015.10.117 (2016).
9. Haque, A., Nguyen, P., Bliek, F. & Slootweg, J. Demand response for real-time congestion management incorporating dynamic thermal overloading cost. *Sustain. Energy Grids Netw.* **10**, 65–74. https://doi.org/10.1016/j.segan.2017.03.002 (2017).
10. Bruninx, K., Dvorkin, Y., Delarue, E., D'haeseleer, W. & Kirschen, D. S. Valuing demand response controllability via chance constrained programming. *IEEE Trans. Sustain. Energy* **9**, 178–187. https://doi.org/10.1109/TSTE.2017.2718735 (2018).
11. Harder, N., Weidlich, A. & Staudt, P. Finding individual strategies for storage units in electricity market models using deep reinforcement learning. *Energy Inform.* **6**, 488–502. https://doi.org/10.1186/s42162-023-00293-0 (2023).
12. Lund, P. D., Lindgren, J., Mikkola, J. & Salpakari, J. Review of energy system flexibility measures to enable high levels of variable renewable electricity. *Renew. Sustain. Energy Rev.* **45**, 785–807. https://doi.org/10.1016/j.rser.2015.01.057 (2015).
13. International Renewable Energy Agency. Power system flexibility for the energy transition, part 1: Overview for policy makers. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2018/Nov/IRENA_Power_system_flexibility_1_2018.pdf (2018).
14. Strbac, G. et al. Role and value of flexibility in facilitating cost-effective energy system decarbonisation. *Prog. Energy* **2**, 042001. https://doi.org/10.1088/2516-1083/abb216 (2020).
15. Dinga, C. D., Van Rijn, S., De Vries, L. & Cvetkovic, M. Coordinated flexibility scheduling in multi-carrier integrated energy systems: A model coupling approach. In *2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*. 1–5. https://doi.org/10.1109/ISGTEUROPE62998.2024.10863244 (2024).
16. Stawska, A., Romero, N., de Weerdt, M. & Verzijlbergh, R. Demand response: For congestion management or for grid balancing?. *Energy Policy* **148**, 111920. https://doi.org/10.1016/j.enpol.2020.111920 (2021).
17. Wang, Y. et al. Operation optimization of regional integrated energy system based on the modeling of electricity-thermal-natural gas network. *Appl. Energy* **251**, 113410. https://doi.org/10.1016/j.apenergy.2019.113410 (2019).
18. Zhang, Y., Li, J., Ji, X., Yang, M. & Ye, P. Optimal scheduling of electricity-gas-heat integrated energy system with coordination of flexibility and reliability. *Sustain. Energy Technol. Assess.* **71**, 103968. https://doi.org/10.1016/j.seta.2024.103968 (2024).
19. Martínez Ceseña, E. A. & Mancarella, P. Energy systems integration in smart districts: Robust optimisation of multi-energy flows in integrated electricity, heat and gas networks. *IEEE Trans. Smart Grid* **10**, 1122–1131. https://doi.org/10.1109/TSG.2018.2828146 (2019).
20. Li, Y., Wang, B., Yang, Z., Li, J. & Li, G. Optimal scheduling of integrated demand response-enabled community-integrated energy systems in uncertain environments. *IEEE Trans. Indus. Appl.* **58**, 2640–2651. https://doi.org/10.1109/TIA.2021.3106573 (2022).
21. Patteeuw, D. et al. Integrated modeling of active demand response with electric heating systems coupled to thermal energy storage systems. *Appl. Energy* **151**, 306–319. https://doi.org/10.1016/j.apenergy.2015.04.014 (2015).
22. Zhang, X., Strbac, G., Shah, N., Teng, F. & Pudjianto, D. Whole-system assessment of the benefits of integrated electricity and heat system. *IEEE Trans. Smart Grid* **10**, 1132–1145. https://doi.org/10.1109/TSG.2018.2871559 (2019).
23. Zhang, X., Shahidehpour, M., Alabdulwahab, A. & Abusorrah, A. Hourly electricity demand response in the stochastic day-ahead scheduling of coordinated electricity and natural gas networks. *IEEE Trans. Power Syst.* **31**, 592–601. https://doi.org/10.1109/TPWRS.2015.2390632 (2016).
24. Gottwalt, S., Gärttner, J., Schmeck, H. & Weinhardt, C. Modeling and valuation of residential demand flexibility for renewable energy integration. *IEEE Trans. Smart Grid* **8**, 2565–2574. https://doi.org/10.1109/TSG.2016.2529424 (2017).
25. Dong, H., Shan, Z., Zhou, J., Xu, C. & Chen, W. Refined modeling and co-optimization of electric-hydrogen-thermal-gas integrated energy system with hybrid energy storage. *Appl. Energy* **351**, 121834. https://doi.org/10.1016/j.apenergy.2023.121834 (2023).
26. Gusain, D. *Flexibility in Multi Energy Systems: Models and Metrics to Assess Future Energy Systems*. PhD Dissertation, Delft University of Technology. https://doi.org/10.4233/uuid:3e857b6c-b028-4f40-9625-54ca79477be8 (2022).
27. Palensky, P., Van Der Meer, A. A., Lopez, C. D., Joseph, A. & Pan, K. Cosimulation of intelligent power systems: Fundamentals, software architecture, numerics, and coupling. *IEEE Indus. Electron. Mag.* **11**, 34–50. https://doi.org/10.1109/MIE.2016.2639825 (2017).
28. Deane, J., Chiodi, A., Gargiulo, M. & Ó Gallachóir, B. P. Soft-linking of a power systems model to an energy systems model. *Energy* **42**, 303–312. https://doi.org/10.1016/j.energy.2012.03.052 (2012) (8th World Energy System Conference, WESC 2010).
29. Schledorn, A., Junker, R. G., Guericke, D., Madsen, H. & Dominkovi, D. F. Frigg: Soft-linking energy system and demand response models. *Appl. Energy* **317**, 119074. https://doi.org/10.1016/j.apenergy.2022.119074 (2022).
30. Zhang, M. et al. An novel soft-linking model: Coordinating the regulative flexibility and operational stability of pumped storage units. *Appl. Energy* **374**, 123942. https://doi.org/10.1016/j.apenergy.2024.123942 (2024).
31. Rajakovic, N., Ivanović, B., Bjelic, I. B. & Rajić, T. Coupling of the planning and power system analysis software for high-level penetration of renewable energy sources. *Energy Convers. Manag. X* **28**, 101298. https://doi.org/10.1016/j.ecmx.2025.101298 (2025).

32. Rosendal, M. et al. The benefits and challenges of soft-linking investment and operational energy system models. *Appl. Energy* **385**, 125512. https://doi.org/10.1016/j.apenergy.2025.125512 (2025).
33. Palensky, P., Mancarella, P., Hardy, T. & Cvetkovic, M. Cosimulating integrated energy systems with heterogeneous digital twins: Matching a connected world. *IEEE Power Energy Mag.* **22**, 52–60. https://doi.org/10.1109/MPE.2023.3324886 (2024).
34. Han, S., Topcu, U. & Pappas, G. J. Differentially private distributed constrained optimization. *IEEE Trans. Autom. Control* **62**, 50–64. https://doi.org/10.1109/TAC.2016.2541298 (2017).
35. Aswani, A., Shen, Z. M. & Siddiq, A. Inverse optimization with noisy data. *Oper. Res.* **66**, 870–892. https://doi.org/10.1287/opre.2017.1705 (2018).
36. Dvorkin, V., Fioretto, F., Van Hentenryck, P., Pinson, P. & Kazempour, J. Differentially private optimal power flow for distribution grids. *IEEE Trans. Power Syst.* **36**, 2186–2196. https://doi.org/10.1109/TPWRS.2020.3031314 (2021).
37. Widl, E. et al. Expert survey and classification of tools for modeling and simulating hybrid energy networks. *Sustain. Energy Grids Netw.* **32**, 100913. https://doi.org/10.1016/j.segan.2022.100913 (2022).
38. Chang, M., Lund, H., Thellufsen, J. Z. & Østergaard, P. A. Perspectives on purpose-driven coupling of energy system models. *Energy* **265**, 126335. https://doi.org/10.1016/j.energy.2022.126335 (2023).
39. Steinbrink, C. et al. CPES testing with mosaik: Co-simulation planning, execution and analysis. *Appl. Sci.* **9**. https://doi.org/10.3390/app9050923 (2019).
40. Gusain, D., Cvetkovi, M. & Palensky, P. Simplifying multi-energy system co-simulations using energysim. *SoftwareX* **18**, 101021. https://doi.org/10.1016/j.softx.2022.101021 (2022).
41. Hardy, T. D., Palmintier, B., Top, P. L., Krishnamurthy, D. & Fuller, J. C. Helics: A co-simulation framework for scalable multi-domain modeling and analysis. *IEEE Access* **12**, 24325–24347. https://doi.org/10.1109/ACCESS.2024.3363615 (2024).
42. Kargarian, A. et al. Toward distributed/decentralized dc optimal power flow implementation in future electric power systems. *IEEE Trans. Smart Grid* **9**, 2574–2594. https://doi.org/10.1109/TSG.2016.2614904 (2018).
43. Vázquez-Canteli, J. R. & Nagy, Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Appl. Energy* **235**, 1072–1089. https://doi.org/10.1016/j.apenergy.2018.11.002 (2019).
44. Saad, W., Han, Z., Poor, H. V. & Basar, T. Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications. *IEEE Signal Process. Mag.* **29**, 86–105. https://doi.org/10.1109/MSP.2012.2186410 (2012).
45. Doh Dinga, C. & van Rijn, S. *Demoses-admm* https://doi.org/10.5281/zenodo.15116423 (2025).
46. Doh Dinga, C. & van Rijn, S. Annular v0.2.0. https://doi.org/10.5281/zenodo.15496488 (2025).
47. Doh Dinga, C. & van Rijn, S. *On the smart coordination of flexibility scheduling in multi-carrier integrated energy systems* https://doi.org/10.5281/zenodo.15496935 (2025).
48. Brown, T., Schlachtberger, D., Kies, A., Schramm, S. & Greiner, M. Synergies of sector coupling and transmission reinforcement in a cost-optimised, highly renewable European energy system. *Energy* **160**, 720–739. https://doi.org/10.1016/j.energy.2018.06.222 (2018).
49. Poplavskaya, K. et al. Integration of day-ahead market and redispatch to increase cross-border exchanges in the European electricity market. *Appl. Energy* **278**, 115669. https://doi.org/10.1016/j.apenergy.2020.115669 (2020).
50. Neumann, F., Zeyen, E., Victoria, M. & Brown, T. The potential role of a hydrogen network in Europe. *Joule* **7**, 1793–1817. https://doi.org/10.1016/j.joule.2023.06.016 (2023).
51. Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* (Now Foundations and Trends, 2011).
52. Bruninx, K., Ovaere, M. & Delarue, E. The long-term impact of the market stability reserve on the EU emission trading system. *Energy Econ.* **89**, 104746. https://doi.org/10.1016/j.eneco.2020.104746 (2020).
53. Höschle, H., Le Cadre, H., Smeers, Y., Papavasiliou, A. & Belmans, R. An admm-based method for computing risk-averse equilibrium in capacity markets. *IEEE Trans. Power Syst.* **33**, 4819–4830. https://doi.org/10.1109/TPWRS.2018.2807738 (2018).
54. Doh Dinga, C. & van Rijn, S. *Cronian*. https://doi.org/10.5281/zenodo.13142375 (2025).
55. SURF. *Snellius, the National Supercomputer*.
56. Gøtske, E. K., Andresen, G. B., Neumann, F. & Victoria, M. Designing a sector-coupled European Energy System robust to 60 years of historical weather data. *Nat. Commun.* **15**, 10680. https://doi.org/10.1038/s41467-024-54853-3 (2024).
57. Glaum, P., Neumann, F., Millinger, M. & Brown, T. A minimal methanol backstop for high-electrification scenarios. *Joule* https://doi.org/10.1016/j.joule.2026.102464 (2026).
58. Xiong, B., Brown, T. & Riepin, I. The role of projects of common interest in reaching Europe's energy policy targets. *Environ. Res. Lett.* **21**, 034011. https://doi.org/10.1088/1748-9326/ae3846 (2026).
59. Neumann, F., Hampp, J. & Brown, T. Green energy and steel imports reduce Europe's net-zero infrastructure needs. *Nat. Commun.* **16**, 5302. https://doi.org/10.1038/s41467-025-60652-1 (2025).
60. Borenstein, S., Bushnell, J. & Wolak, F. A. Measuring market inefficiencies in California's restructured wholesale electricity market. *Am. Econ. Rev.* **92**, 1376–1405. https://doi.org/10.1257/000282802762024557 (2002).
61. Cramton, P. Electricity market design. *Oxf. Rev. Econ. Policy* **33**, 589–612 (2017).
62. Gurobi Optimization, LLC. Managing threads and memory usage in Gurobi. https://support.gurobi.com/hc/en-us/articles/42058344459409. Accessed 29 May 2026 (2025).

Acknowledgements

This publication is part of the DEMOSES project, financed by the Dutch Research Council (NWO) under grant ID ESI.2019.004. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative, using grant no. EINF-11996.

Author contributions

C.D.D. and M.C; conceptualization, C.D.D. and S.V.R; methodology and software, C.D.D.; data curation, investigation and visualization, C.D.D.; writing-–original draft, C.D.D. and and S.V.R; writing-–review & editing, L.J.D.V. and M.C.; validation, L.J.D.V. and M.C.; funding acquisition, L.J.D.V. and M.C.; resources, L.J.D.V. and M.C.; supervision.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.D.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026