Optimization and Engineering manuscript No.
(will be inserted by the editor)

A Constraint Disaggregation Method for Structure-Preserving Aggregations in LP Problems: Application to Renewable Energy Grids with Hydrogen Storage

Gabor Riccardi · Bianca Urso · Stefano Gualandi

Received: date / Accepted: date

Abstract In recent years, the integration of renewable energy sources into electrical grids has become a critical area of research due to the increasing need for sustainable and resilient energy systems. To address the variability of wind and solar power output over time, electricity grids expansion plans need to account for multiple scenarios over large time horizons. This significantly increases the size of the resulting Linear Programming (LP) problem, making it computationally challenging for large scale grids. To tackle this, we propose an approach that aggregates time-steps to reduce the problem size, followed by an iterative refinement of the aggregation, in order to converge to the optimal solution. Using the previous iteration's solution as a warm start, we introduce and compare methods to select which time intervals to refine at each iteration. The first method selects time-steps based on the proportion of net power production within each interval. We provide a general theoretical justification for its use and sufficient conditions under which an optimal solution of an aggregated linear problem can extend to an optimal solution of the original problem. The second method employs a Rolling Horizon (RH) method to evaluate the feasibility of the aggregated solutions, and selects the time interval on which the validation fails. These selection methods are then compared against a random interval selection approach.

Keywords Electric Power System · Stochastic Programming · Rolling Horizon · Time Series Aggregation · Renewable Energy

Mathematics Subject Classification (2020) 90-10 · 90B15

Gabor Riccardi, Stefano Gualandi
Dipartimento di Matematica "F. Casorati" Via Adolfo Ferrata, 5, 27100 Pavia, Italy
E-mail: gabor.riccardi01@universitadipavia.it. E-mail: stefano.gualandi@unipv.it

Bianca Urso
IUSS School of Advances Studies, Palazzo del Broletto, 27100 Pavia, Italy
E-mail: bianca.urso@iusspavia.it

1 Introduction

The threat of climate change is pushing policy-makers to pursue greater integration of renewable energy sources into electrical grids while at the same time ensuring reliability and resilience through digital optimization of electric energy distribution and transmission in smart grids (European Commission 2024). One of the main difficulties arising when designing an electric energy system relying on renewable sources is the great variability of electricity generation through wind and photovoltaic since these resources are highly dependent on weather conditions. To deal with this variability, a possible solution gaining a lot of traction in recent years is the introduction of an energy storage system relying on hydrogen, converting energy from hydrogen to electricity and vice versa in fuel cells and electrolyzers (Blanco and Faaij (2018); Parra et al. (2019)). It is of interest to evaluate the optimal solution, in terms of investment plan, to supply the grid along with industrial hydrogen demand in a dependable way. The stochastic nature of the problem makes it impossible to plan long-term by optimizing using weather forecasts, and a statistical approach is required to ensure a robust model.

Up to now, common approaches have adopted Stochastic Programming or Robust Optimization models, along with hybrid models involving Information Gap Decision Theory or Chance Constraint (Jasiński et al. 2023). While initially favored, the stochastic programming approach comes with a high computational burden, so robust optimization models have seen more popularity in recent years despite the drawback of being conservative methods with a higher average cost of operation and planning of energy systems.

In the typical setting, the problem to solve is a Capacity Expansion Problem (CEP) regarding infrastructure investments: solar and wind farms, fuel cells, hydrolizers, and grid upgrades to augment Net Transfer Capacity (NTC). Nested within the CEP is an Economic Dispatch (ED) problem concerning the operational costs of the infrastructure. The problem is well suited to be modeled using mixed integer linear programming (MILP), as is explained in detail, for example, in Morais et al. (2010).

The CEP for investment planning requires looking at long time horizons. On the other hand, intra-day variability in power generation is the main complexity driver for the ED problem, so the time horizon must be modeled by a large number of fine-grained time-steps. Furthermore, large-scale grids can be modeled with various degrees of spatial aggregation, as is explored in Hörsch and Brown (2017) and Biener and Garcia Rosas (2020), and problem size increases more than linearly with the number of nodes. Thus, the temporal and spatial characteristics of the model cause the MILP size to increase rapidly. This is especially demanding in stochastic programming, since all the variables from the inner ED problem must be reproduced over all scenarios.

To reduce these costs, one possible approach is to use a Rolling Horizon (RH). The basic technique is described in the work of Glomb et al. (2022), along with some results regarding quality guarantees for the optimality of the solution. In Palma-Behnke et al. (2013), a rolling horizon approach is used

within a robust optimization model to optimize the operation of a microgrid composed of two photovoltaic (PV) systems, a wind turbine, a diesel generator, and a lead-acid battery for storage, serving an isolated area in Chile. A similar idea, denoted as the "fix-and-relax method", is applied in the work of Yilmaz et al. (2020), where integer variables representing capital investments are initially relaxed and then progressively fixed in successive time-steps, reducing the computational costs associated with the search for integer solutions. The same method is applied by Kirschbaum et al. (2023) for optimizing medium-scale industrial energy systems.

A severe drawback of the RH approach is that the solution it provides is not optimal over the whole time horizon. Indeed Keppo and Strubegger (2010) explore the effect of short-term planning with limited foresight compared to perfect foresight optimization. On the other hand, the RH approach better reflects actual decision-making based on information that is only progressively available, which is the case for the management of energy system planning relying on weather forecasts.

Methods for representing the temporal dimension to reduce computational costs are broadly categorized as Time Series Aggregation methods. For a comprehensive review of these methods, see Teichgraeber and Brandt (2022). These approaches fall into several categories: *resolution variation methods*, where adjacent time-steps are aggregated into longer time-steps of equal length (downsampling) or into time-steps of varying lengths based on selected temporal features (segmentation), and Multiple Time Grid methods, which consider different timescales for network components operating at different rates.

Various features can guide the aggregation of time series, such as the gradient magnitude between adjacent time-steps Mavrotas et al. (2008) or k-means clustering on heat profiles Fazlollahi et al. (2014). Instead of aggregating time-steps into a single longer time-step, other methods focus on selecting a subset of periods to represent the full series. These periods may be chosen randomly (*Random Sampling*) or by the *Typical Periods* (or *Typical Days*) approach, which selects a representative subset of days for each season.

For example, Domínguez-Muñoz et al. (2011) uses clustering to select representative days before optimization, and Marquant et al. (2017) compares the performance of Typical Days and Rolling Horizon (RH) methods in terms of solving time and accuracy for a selected distributed energy system model.

In our work, we build an LP model of a large-scale electrical grid powered by wind and solar power generation and supported by hydrogen storage. The LP aims to solve the CEP for the grid design, optimizing stochastically on the scenarios for the inner ED problem. We present our approach to efficiently dealing with the computational burden resulting from aggregating time-steps to reduce the model size. The optimization is carried out using an iterative procedure that gradually refines the partition of the time horizon and converges to the optimal solution. Two methods are discussed and compared to guide the selection of these progressively tighter relaxations of the perfect foresight model. The first method selects time-steps based on the proportion of net power production within each interval. We provide a theoretical justification

for using this index for a broader class of Linear Problems, establishing sufficient conditions under which an optimal solution to an aggregated problem can be extended to an optimal solution of the original linear problem. The second method leverages a Rolling Horizon (RH) approach, where we refine the time interval in which the RH fails. This latter method finds feasible solutions within a limited-foresight environment.

Our contributions. The main contributions of this paper are:

1. We introduce the concept of structure-preserving constraint transformations in Linear Programming and establish sufficient conditions for extending the solution of a transformed problem to an optimal solution of the original problem.
2. Building on these results, we identify sufficient conditions under which a solution to a Capacity Expansion Problem with aggregated time series remains optimal for the original problem.
3. We develop two novel heuristics for refining time series aggregations in the Capacity Expansion Problem: one leverages the proposed structure-preserving constraint aggregation framework, and the other employs a rolling horizon approach.
4. We evaluate the performance of these heuristics by comparing them against an iterative method that randomly selects time intervals for refinement.

Outline. The structure of this paper is as follows. Section 2 introduces the Capacity Expansion Problem (CEP) formulation in the context of time series aggregation. Section 3 develops the concept of *structure-preserving* aggregations for linear problems and establishes the conditions under which an aggregated solution can be extended to an optimal solution of the original problem. In Section 4, we apply these results to the aggregated CEP, deriving sufficient conditions for extending an optimal solution of the aggregated problem to the original CEP. Section 5 presents two iterative methods for refining time aggregations to converge toward an optimal perfect foresight solution. The first method leverages a feasibility index introduced in Section 5, while the second employs a rolling horizon approach. Finally, Section 6 discusses the computational results and compares the performance of the proposed methods.

1.1 Notation

We denote a general Linear Programming (LP) problem as

$$\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{x} \in P\}, \text{ with } P := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}, \tag{1}$$

where $\mathbf{A}$ is an $m \times n$ matrix. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with row and column index sets $I \subset \{1,\dots,m\}$ and $J \subset \{1,\dots,n\}$, respectively, we denote the submatrix of $\mathbf{A}$ with rows in I and columns in J as $\mathbf{A}_{I,J}$. Let $\sigma = \{R_1, R_2, \dots, R_{\tilde{m}}\}$ denote a partition of the m rows with $\tilde{m} < m$, and

$\delta = \{C_1, C_2, \ldots, C_{\tilde{n}}\}$ a partition of the n the columns with $\tilde{n} < n$. For a family of sets F, we denote the elements in F with size exactly k and greater than k by $F_{=k}$ and $F_{>k}$, respectively. Then, $\text{supp}(\tilde{\mathbf{A}}_R)_{>1} \subset \delta$ represents the set of indices corresponding to partitions $C \in \delta$ with size greater than 1, and $\text{supp}(\mathbf{A}_r)_{>1}$ refers to the set of indices where $c \in C \in \delta$, with size greater than 1.

Let $\tilde{\mathbf{A}}$ be formed by aggregating the rows and columns of $\mathbf{A}$ according to the partitions σ and δ, respectively. For each $R \in \sigma$, we denote by $\tilde{\mathbf{A}}_R$ the row in $\tilde{\mathbf{A}}$ resulting from aggregating the rows of $\mathbf{A}$ corresponding to R, while $\mathbf{A}_R$ refers to the submatrix of $\mathbf{A}$ consisting of all rows in R. Similarly, for each $C \in \delta$, we define $\tilde{\mathbf{A}}^C$ as the column in $\tilde{\mathbf{A}}$ obtained by aggregating the columns of $\mathbf{A}$ in C, and $\mathbf{A}^C$ as the submatrix of $\mathbf{A}$ containing all columns in C. Thus, σ and δ serve as the index sets for $\tilde{\mathbf{A}}$. For all n in $\mathbb{N}$ with $n \geq 1$, we refer to the set $\{1, \ldots, n\}$ as $[n]$.

2 Problem

This section introduces our LP model for the Capacity Expansion Problem ($\text{CEP}_{\mathcal{T}}$) with perfect foresight over a time horizon $\mathcal{T}$. The model describes a European network that will be powered and supplied with hydrogen using electric power generated by photovoltaic panels and wind turbines, converted to hydrogen through electrolysis, and potentially reconverted in fuel cells.[1]

2.1 Modeling the Capacity Expansion Problem

The European network is represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ corresponds to the nodes (buses) in the network, and $\mathcal{L} = \mathcal{L}_H \cup \mathcal{L}_P$ represents transmission lines ($l \in \mathcal{L}_P$) and hydrogen lines ($l \in \mathcal{L}_H$). Each node has its generators, hydrogen storage, fuel cells, and hydrolyzers. The Net Transfer Capacity (NTC) is defined as the maximum power flow allowed through a transmission line ($l \in \mathcal{L}_P$). The model aims to determine the optimal capacities for all network components, including the Net Transfer Capacity for each transmission line, the maximum transmission capacity for hydrogen pipelines ($l \in \mathcal{L}_H$), and the capacities of generators, hydrogen storage units, fuel cells, and electrolyzers at each node.

The model considers the generation and load scenarios of the given area along with various parameters reflecting the costs and efficiency of the current state of technology and physical upper bounds for the decision variables. When jointly optimizing over multiple scenarios, the solver returns the minimal amount of infrastructure and capacities needed to have a feasible solution,

[1] This work is motivated by the participation of the first two authors to the AIMMS-MOPTA 2024 competition entitled *Would a Fully Renewable Energy Grid benefit from adding Green Hydrogen as a Supplemental Power Source?* (full description at `https://coral.ise.lehigh.edu/~mopta/competition`).

entailing meeting demand at all times without any blackouts, over all scenarios and all time-steps in the time horizon, with minimal cost.

In the following paragraphs, we describe the three main components of our model: the decision variables, the objective function, and the network constraints.

Decision Variables The primary decision variables of interest for policymakers are those linked to infrastructure that characterize the grid design. Table 1 summarizes all decision variables. In particular, for each node $i \in \mathcal{N}$, the policymakers have to decide the number of wind turbines $x_i^{(w)}$ and photovoltaic panels $x_i^{(p)}$ to be installed and the required hydrogen capacity storage $x_i^{(h)}$. In our model, stored hydrogen encompasses liquid and gaseous forms without distinguishing between the two. We assume that hydrogen is immediately available for long-term storage upon conversion from electricity and can be instantaneously converted back into electricity using fuel cells whenever required.

The key decision variables in grid planning include the power capacity level and the conversion speed of fuel cells and electrolyzers. Specifically, the variables $x_i^{(h\to e)}$ and $x_i^{(e\to h)}$, indexed by node $i \in \mathcal{N}$, represent the maximum amount of energy that can be converted at a single time-step from hydrogen to electric energy and vice versa, respectively. Accurately planning these quantities is crucial for designing a grid capable of effectively dealing with peak production and ensuring supply during periods of low production. Net Transfer Capacity constrains transmission across the power grid, while hydrogen transfer is constrained by pipeline capacity. Potential improvements to the existing power lines or hydrogen transport infrastructure capacity are represented by the decision variables $y_l^{(e)}$ and $y_l^{(h)}$.

In our problem, we consider a time horizon $\mathcal{T} = \{1, \dots, T\}$, and we optimize over a collection of scenarios $\mathcal{J} = \{1, \dots, J\}$. The variables independent from the scenario are called *first-stage* variables. Instead, the variables indexed by a scenario $s \in \mathcal{J}$ and a time-step $t \in \mathcal{T}$ are called *second-stage* variables.

Table 2 describes all the costs and parameters used in our model, which we will detail in the following paragraphs.

Objective Function. The objective function is the sum of all capital costs associated with infrastructure installation, combined with the marginal costs for hydrogen-to-electricity $(h \to e)$ and electricity-to-hydrogen $(e \to h)$ conversions, and the hydrogen transfer costs across the lines at every time-step, averaged over the given collection of scenarios. Overall, the objective function is as follows:

Table 1 Summary of decision variables.

Name	Unit	Description
$x_i^{(p)}$	-	Photovoltaic solar panels at node i
$x_i^{(w)}$	-	Wind turbines at node i
$x_i^{(h)}$	kg	Hydrogen storage capacity at node i
$x_i^{(h\to e)}$	kg	Hydrogen to electricity conversion capacity per time-step at node i
$x_i^{(e\to h)}$	MWh	Electricity to hydrogen conversion capacity per time-step at node i
$y_l^{(e)}$	MWh	Additional net transfer capacity on line $l \in \mathcal{L}_P$
$y_l^{(h)}$	kg	Additional hydrogen transfer capacity on pipe $l \in \mathcal{L}_H$
z_{ist}^{Δ}	kg	Hydrogen extracted (or injected) from storage at node i, scenario s, time t
z_{ist}	kg	Total Hydrogen stored at node i, scenario s, time t
$z_{ist}^{(h\to e)}$	kg	Hydrogen to electricity in node i, scenario s, time t
$z_{ist}^{(e\to h)}$	MWh	Electricity to hydrogen in node i, scenario s, time t
$z_{lst}^{(e)}$	MWh	Electricity passing through line l, scenario s, time t
$z_{lst}^{(h)}$	kg	Hydrogen transported on pipe l, scenario s, time t

$$
\begin{aligned}
\min \sum_{i\in\mathcal{N}} \Big(& p_i x_i^{(p)} + w_i x_i^{(w)} + h_i x_i^{(h)} + c_i^{(h\to e)} x_i^{(h\to e)} + c_i^{(e\to h)} x_i^{(e\to h)} \Big) + \\
& \sum_{l\in\mathcal{L}_P} d_l^{(e)} y_l^{(e)} + \sum_{l\in\mathcal{L}_H} d_l^{(h)} y_l^{(h)} + \\
& \frac{1}{J} \sum_{s\in\mathcal{J}} \sum_{t\in\mathcal{T}} \Bigg(\sum_{i\in\mathcal{N}} \Big(c_i^{(h\,\to\, e)} z_{ist}^{(h\to e)} + c_i^{(e\,\to\, h)} z_{ist}^{(e\to h)} \Big) + \\
& \qquad \sum_{l\in\mathcal{L}_H} c_l^{(h)} |z_{lst}^{(h)}| + \sum_{l\in\mathcal{L}_P} c_l^{(e)} |z_{lst}^{(e)}| \Bigg),
\end{aligned}
\tag{2}
$$

where the $\frac{1}{J}$ factor applied to the marginal costs enables averaging across scenarios, while the capital costs remain constant for all scenarios. Excluding the costs of $x_i^{(h\to e)}$ and $x_i^{(e\to h)}$, the objective function provides an estimate of the actual costs (in euros) associated with the system's setup and maintenance over the specified time horizon.

Network Constraints. The following constraints ensure that electricity and hydrogen loads are met for each time-step t and scenario s. Let $\delta^+(i)$ and $\delta^-(i)$ indicate the outgoing and incoming edges from node $i \in \mathcal{N}$. Then, the following flow balance constraints are imposed:

Table 2 Model costs and parameters.

Name	Unit	Description
p_i	€	Unitary cost of a Photovoltaic Panel
w_i	€	Unitary cost of a Wind Turbine
h_i	€/kg	Cost of hydrogen storage capacity
$c_i^{(h\to e)}$	€/kg	Hydrogen to electricity conversion cost
$b_i^{(h\to e)}$	€/kg	Hydrogen to electricity unitary capacity cost
$f_i^{(h\to e)}$	-	Hydrogen to electricity conversion efficiency
$c_i^{(e\to h)}$	€/MWh	Electricity to hydrogen conversion cost
$b_i^{(e\to h)}$	€/MWh	Electricity to hydrogen unitary capacity cost
$f_i^{(e\to h)}$	-	Electricity to hydrogen conversion efficiency
$c_l^{(h)}$	€/kg	Cost of transferring hydrogen on pipe $l \in \mathcal{L}_H$
$c_l^{(e)}$	€/MWh	Cost of transferring power on line $l \in \mathcal{L}_P$
$d_l^{(e)}$	€/MWh	Cost of adding net transfer capacity to line $l \in \mathcal{L}_P$
$d_l^{(h)}$	€/kg	Cost of adding hydrogen transfer capacity to pipe $l \in \mathcal{L}_H$
$a_l^{(e)}$	MWh	Existing net transfer capacity on line $l \in \mathcal{L}_P$
$a_l^{(h)}$	kg	Existing hydrogen transfer capacity on pipe $l \in \mathcal{L}_H$
$E_{ist}^{(e)}$	MWh	Electric energy output of a single photovoltaic panel
$W_{ist}^{(e)}$	MWh	Electric energy output of a single wind turbine
$L_{ist}^{(e)}$	MWh	Electric energy load (per node i, scenario s, time-step t)
$L_{ist}^{(h)}$	kg	Hydrogen load (per node i, scenario s, time-step t)

$$\text{Electricity Balance: } 0 = E_{ist}^{(e)} x_i^{(p)} + W_{ist}^{(e)} x_i^{(w)} - L_{ist}^{(e)} + f_i^{(h\to e)} z_{ist}^{(h\to e)} - z_{ist}^{(e\to h)} + \sum_{l\in\delta^+(i)} z_{lst}^{(e)} + \sum_{l\in\delta^-(n)} z_{lst}^{(e)} - s_{ist}, \tag{3}$$

$$\text{Hydrogen Balance: } z_{ist}^{\Delta} = -L_{ist}^{(h)} + f_i^{(e\to h)} z_{ist}^{(e\to h)} - z_{ist}^{(h\to e)} - \sum_{l\in\delta^+(i)} z_{lst}^{(h)} + \sum_{l\in\delta^-(i)} z_{lst}^{(h)}, \tag{4}$$

for each node $i \in \mathcal{N}$, scenario $s \in \mathcal{J}$, and time-step $t \in \mathcal{T}$,

We require that the electricity consumed does not exceed the electricity produced or received at any time. This is modeled using an additional slack variable $s_{ist} \geq 0$. On the grid itself, the two sides should be equal, but we observe that $E_{ist}^{(e)} x_i^{(p)} + W_{ist}^{(e)} x_i^{(w)}$ indicate the maximum power that can be

generated with given weather conditions, whereas the actual production will be regulated to meet demand through curtailment.

We link the variables corresponding to hydrogen extracted from storage and hydrogen stored at each time-step through the following constraints.

$$\text{Hydrogen storage: } z_{is,\bar{t}+1} = z_{is,1} + \sum_{t=1}^{\bar{t}} z^{\Delta}_{ist},\ \bar{t} = 1, .., T-1, \forall i \in \mathcal{N}, \forall s \in \mathcal{J}, \quad (5)$$

where T is the last time-step in the time horizon $\mathcal{T}$.

To prevent the optimal solution from taking artificially high initial hydrogen storage levels, we force hydrogen storage at the last time-step to be equal to the initial hydrogen storage using the following constraint:

$$\text{Periodic Hydrogen Storage:} \quad \sum_{t \in \mathcal{T}} z^{\Delta}_{ist} = 0, \quad \forall i \in \mathcal{N}, \forall s \in \mathcal{J}. \quad (6)$$

The total storage and conversion capacities are calculated by minimizing the maximum over time and scenarios of the respective variables:

$$\text{Storage Capacity Limit:} \quad z_{ist} \leq x^{(h)}_i, \qquad \forall i \in \mathcal{N}, \forall s \in \mathcal{J}, \forall t \in \mathcal{T}, \quad (7)$$

$$\text{EtH Conversion Limit:} \quad z^{(e \to h)}_{ist} \leq x^{(e \to h)}_i, \quad \forall i \in \mathcal{N}, \forall s \in \mathcal{J}, \forall t \in \mathcal{T}, \quad (8)$$

$$\text{HtE Conversion Limit:} \quad z^{(h \to e)}_{ist} \leq x^{(h \to e)}_i, \quad \forall i \in \mathcal{N}, \forall s \in \mathcal{J}, \forall t \in \mathcal{T}. \quad (9)$$

Finally, we consider edge capacities for net and hydrogen transfer capacity as follows:

$$\text{Net Transfer Capacity:} |z^{(e)}_{lst}| \leq a^{(e)}_l + y^{(e)}_l, \forall l \in \mathcal{L}, \forall s \in \mathcal{J}, \forall t \in \mathcal{T}, \quad (10)$$

$$\text{Hydrogen Transfer Capacity:} |z^{(h)}_{lst}| \leq a^{(h)}_l + y^{(h)}_l, \forall l \in \mathcal{L}, \forall s \in \mathcal{J}, \forall t \in \mathcal{T}. \quad (11)$$

The constraints outlined above set the upper bounds for storage, conversion, and transfer capacities within the unaggregated time horizon $\mathcal{T}$. In the next subsection, we reformulate the Capacity Expansion Problem by substituting the time horizon $\mathcal{T}$ with an aggregated time horizon $\mathcal{P}$, obtained by merging adjacent time-steps in $\mathcal{T}$ into longer intervals.

2.2 Time Series Aggregation

The scenarios generated from our datasets have a time resolution of one hour. Such resolution is enough to capture the daily variability of power generation and load. However, the number of variables and constraints grows linearly with the number of time-steps, rendering the model intractable with just a few scenarios. Moreover, when optimizing over a full year, considering every hour of every day is partly redundant, as each day tends to resemble its neighboring days. On the other hand, simply considering a sample of days for each season might compromise long-term storage capacity representation. Therefore, we

seek more efficient strategies to deal with the time dimension of our problem. We refer to a Time Series Aggregation obtained by joining adjacent time-steps as a *Time Partition*.

Definition 2.1 (Time partition) Given an initial time horizon $\mathcal{T}$, a time partition $P = \{I_1, ..., I_{T'}\}$ is a partition of $\mathcal{T}$ such that all the subsets are time intervals. Furthermore, we say that a time partition $\mathcal{P}'$ is finer than $\mathcal{P}$ if for every $I' \in \mathcal{P}'$, there exists some $I \in \mathcal{P}$ such that $I' \subset I$.

Given a time partition $\mathcal{P}$, we define the problem $\text{CEP}_{\mathcal{P}}$ associated with the Capacity Expansion Problem obtained by considering each interval in $\mathcal{P}$ as a single time-step. The aggregated problem $\text{CEP}_{\mathcal{P}}$ can also be obtained by a row and column aggregation of $\text{CEP}_{\mathcal{T}}$. First, we remove all rows that constrain storage 5 for time-steps within the aggregated periods while retaining the rows corresponding to the end of the aggregated time intervals. This process results in a relaxation of the model. Then, consider the various types of constraints: Electricity Balance (3), Hydrogen Balance (4'), maximum capacity of the time-dependent variables (8)–(11), and bounds on the CEP variables. These constraints are all indexed by a time-step $t \in \mathcal{T}$. To perform the row aggregation, for all time intervals I in $\mathcal{P}$, we sum together the constraints of the same type over the time index t with $t \in I$. Then, we perform column aggregation by substituting the variables corresponding to the following sums with one single aggregated variable:

$$z_{isI}^{(e\to h)} = \sum_{t\in I} z_{ist}^{(e\to h)}, \qquad z_{isI}^{(h\to e)} = \sum_{i\in I} z_{ist}^{(h\to e)}. \tag{12}$$

Similarly, we sum over z_{ist}^{Δ} to obtain the aggregated variable z_{isI}^{Δ}, and we get $z_{slI}^{(e)}$ and $z_{slI}^{(h)}$ by summing over $z_{lst}^{(e)}$ and $z_{lst}^{(h)}$ respectively. Lastly, the cost of the aggregated variables is set equal to the cost of the corresponding unaggregated variables.

3 Structure Preserving Constraint Transformations

The core of our idea is that the constraints transformation done to define $\text{CEP}_{\mathcal{P}}$ "preserves the structure" of the original problem $\text{CEP}_{\mathcal{T}}$. In this section, we formalize this notion of *structure-preserving* transformations for a general Linear Problem and obtain sufficient conditions for an optimal solution of the transformed problem to be extended to an optimal solution for the original LP. Given a row partition $\sigma = \{R_1, \ldots, R_{\tilde{m}}\}$ and a column partition $\delta = \{C_1, \ldots, C_{\tilde{m}}\}$, we can obtain an aggregated problem by replacing each set R_i in σ with a single row, and each set C_j in δ with a single column. An approach to aggregate a set of rows (columns) is by taking a linear combination of the rows (columns). This is known as *weighted aggregation* (e.g., see Zipkin (1977), Zipkin (1980)). Let $\boldsymbol{\omega}_r$ and $\boldsymbol{\tau}_c$ denote the aggregation weight for row

r in $R \in \sigma$ and for column c in $C \in \delta$, respectively. Given a general LP as in 1, the corresponding aggregated LP problem becomes:

$$\min \ \tilde{\mathbf{c}}^T \tilde{\mathbf{x}} \tag{13}$$

$$\text{s.t.} \ \tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} \tag{14}$$

$$\tilde{\mathbf{x}} \geq 0, \tag{15}$$

where $\tilde{\mathbf{A}}$ is a $\tilde{m} \times \tilde{n}$ matrix and $\tilde{\mathbf{x}} \in \mathbb{R}^{\tilde{n}}$. For all R_i in σ and all C_j in δ, the coefficients in $\mathbf{A}_{R_i}^{C_j}$ are substituted by one single coefficient in the matrix $\tilde{\mathbf{A}}$ equal to $\tilde{\mathbf{A}}_{R_i}^{C_j} = \tilde{\mathbf{A}}_i^j \coloneqq \sum_{c \in C_j} \boldsymbol{\tau}_c \sum_{r \in R_i} \boldsymbol{\omega}_r \mathbf{A}_r^c$.

Definition 3.1 Given an LP (1), a row and column aggregation with respect to partitions σ, δ is said to be *structure-preserving* if $f : [n] \to [\tilde{n}]$, given by

$$f : c \mapsto C \quad \text{where } C \text{ is the element in } \delta \text{ such that } c \in C,$$

is such that for each $R \in \sigma_{>1}$ and all $r \in R$:

$$f|_{\text{supp}(\mathbf{A}_r)} : \text{supp}(\mathbf{A}_r) \to \text{supp}(\tilde{\mathbf{A}}_R)$$

is a bijection, and

$$\mathbf{A}_{r,c} = \tilde{\mathbf{A}}_{R,f(c)} \quad \text{for all } c \in \text{supp}(\mathbf{A}_r)_{>1}.$$

This definition implies that the coefficients of the aggregated variables in the aggregated problem match those in the original problem for the corresponding unaggregated variables. Indeed, f can be seen as a function mapping the unaggregated variables to variables of the same "type" in the aggregated constraints.

Observation 3.1 *For all* $\{c\} \in \delta_{=1}$ *we have*

$$f(c) = \{c\}.$$

Example 1 Consider the Capacity Expansion problem for a single scenario of a fully renewable electrical grid as defined in Section 2. The network is represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ corresponds to the nodes (buses) in the network, and $\mathcal{L} = \mathcal{L}_H \cup \mathcal{L}_P$ represents transmission lines ($e \in \mathcal{L}_P$) and hydrogen lines ($e \in \mathcal{L}_H$). At each node i in the network, $x_i^{(p)}$ solar panels and $x_i^{(w)}$ wind turbines are installed, where $x_i^{(p)}$ and $x_i^{(w)}$ are decision variables of the problem. At each time-step $t \in \mathcal{T}$, each solar panel and wind turbine produces $E_{ist}^{(e)}$ MWh and $W_{ist}^{(e)}$ MWh respectively. Furthermore, at every node n, we have power cells allowing for conversion between electricity and hydrogen, modeled through the variables $z_{ist}^{(e \to h)}$ and $z_{ist}^{(h \to e)}$. The power flow through line $l \in \mathcal{L}$ is denoted by the variable $z_{lst}^{(e)}$. At each time-step t

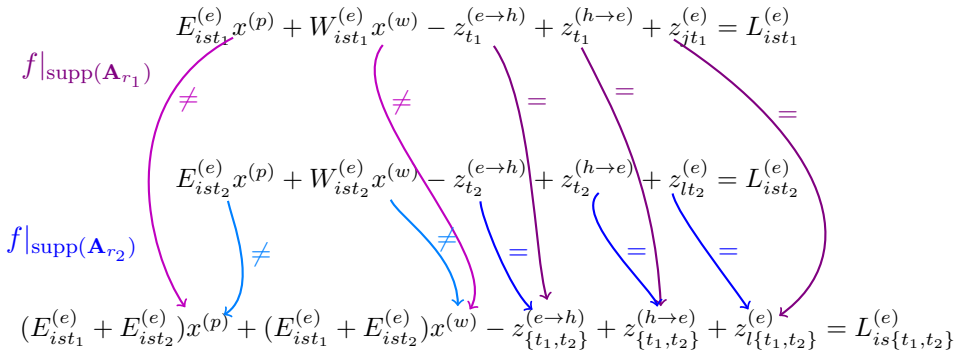

Fig. 1 Example of structure-preserving aggregation

and at each node $i \in \mathcal{N}$, the power injected into node i must equal the power demand $L^{(e)}_{ist\,i,s,t}$, resulting in the power balance constraint:

$$E^{(e)}_{ist\,i,s,t}x^{(p)} + W^{(e)}_{ist\,i,s,t}x^{(w)}_i - z^{(e\to h)}_{ist} + z^{(h\to e)}_{ist} + \sum_{l\in\delta^-(i)} z^{(e)}_{lst} - \sum_{l\in\delta^+(i)} z^{(e)}_{lst} = L^{(e)}_{ist\,i,s,t}$$

Let us examine the Power Balance constraints (3) r_1 and r_2 at a fixed node $i \in \mathcal{N}$ and time-steps t_1 and t_2, respectively. We also define the Power Balance constraint R over the whole time interval $I = \{t_1, t_2\}$: the power produced by each generator equals the sum of power produced during t_1 and t_2, analogously the same is done for the load. For simplicity, we assume that i is connected to a single other node, and we drop index i from the variables. We observe that substituting the constraints r_1 and r_2 with the constraint R is a structure-preserving aggregation with respect to the function f, as depicted in Figure 3, since all the aggregated variables are mapped to variables with the same coefficients. The same can be done for all other constraints appearing in $\mathrm{CEP}_{\mathcal{T}}$, except for the Hydrogen Storage constraint (5), for time-steps $\bar{t}$ where $\bar{t}$ is not an extreme of one of the intervals in $\mathcal{P}$, we refer to these constraints as *intermediate Hydrogen Storage Constraints*. Thus, $\mathrm{CEP}_{\mathcal{P}}$ is a structure-preserving aggregation of $\mathrm{CEP}_{\mathcal{T}}$ without the intermediate Hydrogen Storage Constraints.

While obtaining a feasible solution to problem (1) from the aggregated problem (13) is not always guaranteed, it is possible under certain assumptions.

Observation 3.2 *Let (σ, δ) be a structure-preserving aggregation, $R \in \sigma_{>1}$, and $r \in R$. Let $\tilde{\mathbf{x}}$ be a solution to the aggregated problem (13). If $\tilde{\mathbf{b}}_r - \tilde{\mathbf{A}}_{R,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}} \neq 0$, define*

$$\boldsymbol{\rho}_r := \frac{\mathbf{b}_r - \mathbf{A}_{r,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}}}{\tilde{\mathbf{b}}_r - \tilde{\mathbf{A}}_{R,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}}}.$$

If $\mathbf{A}_{r,\delta_{=1}} = 0$ *and* $\mathbf{b}_r = 0$ *for all* $r \in R$*, then* $\boldsymbol{\rho}_r$ *can be chosen arbitrarily. Otherwise, if* $\boldsymbol{\rho}_r \geq 0$ *and* $\mathbf{x} \in \mathbb{R}^n$ *satisfies* $\mathbf{x}_{\delta_{=1}} = \tilde{\mathbf{x}}_{\delta_{=1}}$ *and* $\mathbf{x}_c = \boldsymbol{\rho}_r \tilde{\mathbf{x}}_{f(C)}$ *for all* $c \in \operatorname{supp}(\mathbf{A}_r)_{>1}$*, then* $\mathbf{x}$ *satisfies the constraints* $\mathbf{A}_r\mathbf{x} = \mathbf{b}_r$ *and* $\mathbf{x}_{\operatorname{supp}(\mathbf{A}_r)} \geq 0$ *of the original problem.*

Proof Consider $\mathbf{A}_r\mathbf{x} = \sum_{i \in \operatorname{supp}(\mathbf{A}_r)} \mathbf{A}_{r,i}\mathbf{x}_i$. This sum can be divided over the unaggregated and aggregated variables:

$$\mathbf{A}_r\mathbf{x} = \mathbf{A}_{r,\delta_{=1}}\mathbf{x}_{\delta_{=1}} + \sum_{c \in \operatorname{supp}(\mathbf{A}_r)_{>1}} \mathbf{A}_{r,c}\mathbf{x}_c. \tag{16}$$

If $\mathbf{A}_{r,\delta_{=1}} = 0$ and $\mathbf{b}_r = 0$ then $\mathbf{A}_{r,\delta_{=1}}\mathbf{x}_{\delta_{=1}} = 0$. Fix $\boldsymbol{\rho}_r \geq 0$, then from the definition of structure-preserving aggregation, we know that $f(\operatorname{supp}(\mathbf{A}_R)_{>1}) = \operatorname{supp}(\tilde{\mathbf{A}}_r)_{>1}$ and $\mathbf{A}_{r,c} = \tilde{\mathbf{A}}_{r,f(c)}$, so (16) becomes:

$$\mathbf{A}_r\mathbf{x} = \sum_{c \in \operatorname{supp}(\mathbf{A}_r)_{>1}} \tilde{\mathbf{A}}_{R,f(c)}\boldsymbol{\rho}_r\tilde{\mathbf{x}}_{f(c)} = \sum_{C \in \operatorname{supp}(\tilde{\mathbf{A}}_R)_{>1}} \tilde{\mathbf{A}}_{R,C}\boldsymbol{\rho}_r\tilde{\mathbf{x}}_C = \boldsymbol{\rho}_r\tilde{\mathbf{A}}_R\tilde{\mathbf{x}} = 0. \tag{17}$$

Thus, $\mathbf{x}$ satisfies the constraint $\mathbf{A}_r\mathbf{x} = \mathbf{b}_r$. When $\mathbf{A}_{r,\delta_{=1}} \neq 0$ or $\mathbf{b}_r \neq 0$, we proceed similarly:

$$\mathbf{A}_r\mathbf{x} = \mathbf{A}_{r,\delta_{=1}}\mathbf{x}_{\delta_{=1}} + \sum_{c \in \operatorname{supp}(\mathbf{A}_r)_{>1}} \tilde{\mathbf{A}}_{R,f(c)}\mathbf{x}_c \tag{18}$$

$$= \mathbf{A}_{r,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}} + \boldsymbol{\rho}_r \sum_{C \in \operatorname{supp}(\tilde{\mathbf{A}}_R)_{>1}} \tilde{\mathbf{A}}_{R,C}\tilde{\mathbf{x}}_C. \tag{19}$$

By the definition of $\boldsymbol{\rho}_r$, in equation (19), the term $\boldsymbol{\rho}_r \sum_{C \in \operatorname{supp}(\tilde{\mathbf{A}}_R)_{>1}} \tilde{\mathbf{A}}_{R,C}\tilde{\mathbf{x}}_C$ is equal to:

$$\begin{aligned}\boldsymbol{\rho}_r \sum_{C \in \operatorname{supp}(\tilde{\mathbf{A}}_R)_{>1}} \tilde{\mathbf{A}}_{R,C}\tilde{\mathbf{x}}_C &= \boldsymbol{\rho}_r(\tilde{\mathbf{A}}_R\tilde{\mathbf{x}} - \tilde{\mathbf{A}}_{R,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}}) \\ &= \boldsymbol{\rho}_r(\tilde{\mathbf{b}}_R - \tilde{\mathbf{A}}_{R,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}}) \\ &= \mathbf{b}_r - \mathbf{A}_{r,\delta_{=1}}\tilde{\mathbf{x}}_{\delta_{=1}}.\end{aligned}$$

Thus, we obtain:

$$\mathbf{A}_r\mathbf{x} = \mathbf{b}_r.$$

□

A structure-preserving aggregation does not inherently ensure the feasibility of all constraints in the original problem. However, Observation 3.2 demonstrates how to partially reconstruct a solution $\mathbf{x}$ for a specific constraint r by scaling the aggregated variables appropriately within the support of $\mathbf{A}_r$.

Definition 3.2 Let $\boldsymbol{\rho}_r$ be defined as in Observation 3.2 for all $r \in R \in \sigma_{>1}$. Let $\mathbf{x} \in \mathbb{R}^n$ be defined as $\mathbf{x}_{\delta_{=1}} \coloneqq \tilde{\mathbf{x}}_{\delta_{=1}}$ and $\mathbf{x}_c \coloneqq \boldsymbol{\rho}_r\tilde{\mathbf{x}}_{f(c)}$ for all $r \in R \in \sigma_{>1}$ and $c \in \operatorname{supp}(\mathbf{A}_r)$. Then, $\mathbf{x}$ is well defined if and only if for all $r, r' \in R \in \sigma_{>1}$ such that $\operatorname{supp}(\mathbf{A}_r) \cap \operatorname{supp}(\mathbf{A}_{r'}) \neq 0$, we have $\boldsymbol{\rho}_r = \boldsymbol{\rho}_{r'}$. If $\boldsymbol{\rho} \geq 0$, we refer to $\mathbf{x}$ as a $\boldsymbol{\rho}$-solution.

If $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution, then $\mathbf{x}$ is feasible for the constraints in $\sigma_{>1}$. In general, $\mathbf{x}$ can violate some of the constraints in $\sigma_{=1}$. However, Observation 3.3 allows us to define a class of constraints in $\sigma_{=1}$ that are always satisfied by a $\boldsymbol{\rho}$-solution.

Observation 3.3 *Let $\boldsymbol{\omega}_r \in \mathbb{R}$ for all $r \in R \in \sigma$ be the weights of the row aggregation. Let $\boldsymbol{\rho}_r$ be defined as in 3.2 for all $r \in R \in \sigma_{>1}$. Then, for all $R \in \sigma_{>1}$ we have:*

$$\boldsymbol{\omega}_R^T \boldsymbol{\rho}_R = 1. \tag{20}$$

Proof (Sketch) By a direct summation, we have

$$\boldsymbol{\omega}_R^T \boldsymbol{\rho}_R = \sum_{r \in R} \boldsymbol{\omega}_r \boldsymbol{\rho}_r = \frac{\sum_{r \in R} \boldsymbol{\omega}_r (\mathbf{b}_r - \mathbf{A}_{r,\delta_{=1}} \tilde{\mathbf{x}}_{\delta_{=1}})}{\tilde{\mathbf{b}}_r - \tilde{\mathbf{A}}_{R,\delta_{=1}} \tilde{\mathbf{x}}_{\delta_{=1}}} = 1.$$

□

Example 2 (Cont. Example 1) In the Capacity Expansion Problem, we fix the storage of hydrogen z_0 at the first time-step at the node i, to be equal to the storage of hydrogen z_T at the last time-step. If we consider the variable $z_{t_i}^{\Delta} \coloneqq z_{t_i} - z_{t_{i-1}}$, this constraint can be expressed as:

$$\sum_{t \in \mathcal{T}}^{T} z_t^{\Delta} = 0. \tag{21}$$

Instead of considering as time-steps $\mathcal{T} = \{0, \dots, T\}$, given a partition $\mathcal{P}$ of $\mathcal{T}$ composed by time intervals, we can instead consider each interval in $\mathcal{P}$ as a single time-step. Then, constraint (21) for a time partition $\mathcal{P}$ becomes:

$$\sum_{I \in \mathcal{P}} \Delta \tilde{z}_I^{\Delta} = 0. \tag{22}$$

Since the weights for the row aggregation are all equal to 1, for Observation 3.3, we have that $\sum_{t \in I} \boldsymbol{\rho}_t = 1$ for all I in $\mathcal{P}$. Given a feasible solution for the aggregated problem $\tilde{z}_I^{\Delta}$, let $z_t^{\Delta} \coloneqq \boldsymbol{\rho}_t \tilde{z}_I^{\Delta}$, then Constraint (21) holds:

$$\sum_{t \in \mathcal{T}} z_t^{\Delta} = \sum_{I \in \mathcal{P}} \sum_{t \in I} \boldsymbol{\rho}_t \tilde{z}_I^{\Delta} = \sum_{I \in P} \tilde{z}_I^{\Delta} = 0 \tag{23}$$

Thus, constraint (21) holds for $\boldsymbol{\rho}$-solutions.

This example is a special instance of a general class of constraints that always hold for $\boldsymbol{\rho}$-solutions. When a $\boldsymbol{\rho}$-solution is well-defined, for each $c \in C \in \delta_{>1}$, we can always select $r_c \in R \in \sigma_{>1}$ such that $\mathbf{x}_c = \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_{f(c)}$. However, a stronger assumption is required: it must be possible to choose all r_c within a single partition set $R^{(C)}$:

Assumption 1 *For each $C \in \delta_{>1}$, there exists distinct $r_c \in [m]$ for all $c \in C$, such that $\mathbf{x}_c = \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_C$ and $R^{(C)} \coloneqq \{r_c\}_{c \in C} \in \sigma_{>1}$.*

Herein, we assume that assumption 1 holds. We observe that this is the case if, for every aggregated variable C, there is an aggregated constraint R for which every element in C appears in distinct rows of R and $|R| = |C|$.

Example 3 (Cont. Example 1) We observe that this assumption holds for the aggregation $\text{CEP}_{\mathcal{P}}$ of $\text{CEP}_{\mathcal{T}}$. Consider the aggregated variables $C \coloneqq \{z_{ist}^{(e\to h)}\}_{t\in I}$ for $I \in \mathcal{T}_{>1}$. Then, we can select $R^{(C)}$ as the set $\{r_{ist}^{E}\}_{t\in I}$ of Electricity Balance constraints (3) for time-steps $t \in I$. Since $R^{(C)} \coloneqq \{r_{ist}^{E}\}_{t\in I}$is in $\sigma_{>1}$, $R^{(C)}$ is a valid choice for Assumption 1.

Assumption 1 holds for C, since if $x = \{z_{ist}^{(e\to h)}\}_{i\in\mathcal{N},s\in\mathcal{J},t\in\mathcal{T}}$ is a $\boldsymbol{\rho}$-solution obtained from the aggregated solution $\tilde{\mathbf{x}} = \{\tilde{z}_{isI}^{(e\to h)}\}_{i\in\mathcal{N},s\in\mathcal{J},I\in\mathcal{P}}$, then, by Observation 3.2, we have $z_{ist}^{(e\to h)} = \boldsymbol{\rho}_{r_{ist}^{E}}\tilde{z}_{isT}^{(e\to h)}$. The same reasoning can be applied to all variables in $\text{CEP}_{\mathcal{P}}$, thus Assumption 1 holds for $\text{CEP}_{\mathcal{P}}$.

Observation 3.4 *Le (σ,δ) be a structure-preserving, row and column aggregation. If $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution and r is a constraint in $\sigma_{=1}$, such that*

$$\mathbf{A}_{r,c} = \boldsymbol{\omega}_{r_c}\tilde{\mathbf{A}}_{r,f(c)} \text{ for all } r_c \in R^{(C)},\ C \in \delta_{>1},$$

where $\boldsymbol{\omega}_{r_c}$ is the aggregation weight for the row r_c, then $\mathbf{x}$ is a feasible solution for constraint r.

Proof As before, we split the sum $\mathbf{A}_r x$ over aggregated and unaggregated variables:

$$\mathbf{A}_r x = \sum_{C\in\delta_{=1}} \mathbf{A}_{r,C}\mathbf{x}_C + \sum_{C\in\delta_{>1}}\sum_{c\in C}\mathbf{A}_{r,c}\mathbf{x}_c. \tag{24}$$

Since r in $\sigma_{=1}$ is not an aggregated constraint, we have that $\mathbf{A}_{r,C} = \tilde{\mathbf{A}}_{r,C}$ and $\mathbf{b}_r = \tilde{\mathbf{b}}_r$. From the hypothesis and the definition of $\boldsymbol{\rho}$-solution, we then have:

$$(24) = \sum_{C\in\delta_{=1}} \tilde{\mathbf{A}}_{r,C}\tilde{\mathbf{x}}_{r,C} + \sum_{C\in\delta_{>1}}\sum_{c\in C}\boldsymbol{\omega}_{r_c}\tilde{\mathbf{A}}_{r,f(c)}\boldsymbol{\rho}_{r_c}\tilde{\mathbf{x}}_{f(c)}. \tag{25}$$

Since $f(c) = C$ for all $c \in C$ and $\boldsymbol{\omega}_{R^{(C)}}\boldsymbol{\rho}_{R^{(C)}} = 1$, we have

$$(25) = \sum_{C\in\delta_{=1}} \tilde{\mathbf{A}}_{r,C}\tilde{\mathbf{x}}_{r,C} + \sum_{C\in\text{supp}(\mathbf{A}_r)_{>1}} \tilde{\mathbf{A}}_{r,C}\tilde{\mathbf{x}}_C = \tilde{\mathbf{A}}_r\tilde{\mathbf{x}}_r = \tilde{\mathbf{b}}_r = \mathbf{b}_r. \tag{26}$$

□

We now define the hypergraph associated with the aggregation (σ,δ).

Definition 3.3 The *hypergraph associated to the aggregation* (σ,δ) is the hypergraph $\mathcal{G} = (\mathcal{N},\mathcal{E})$ having as nodes the unaggregated variables $\mathcal{N} \coloneqq \cup_{C\in\delta_{>1}}C$ and as edges the subsets of $\mathcal{N}$ that appear together in constraints in $\sigma_{>1}$.

When two edges (constraints) in the hypergraph, r and r', share variables, the scaling factors $\boldsymbol{\rho}_r$ and $\boldsymbol{\rho}_{r'}$ must be equal for Observation 3.2 to hold for both r and r'. From this follows the following:

Proposition 3.1 *If (σ, δ) is a structure-preserving aggregation and the constraints in $\sigma_{=1}$ hold for $\boldsymbol{\rho}$-solutions. Let $\tilde{\mathbf{x}}$ be a solution to the aggregated problem* (13)*. For all $r \in R \in \sigma_{>1}$ define $\boldsymbol{\rho}_r$ as in Observation 3.2. If $\boldsymbol{\rho}_r \geq 0$ and is constant over the connected components of the hypergraph associated to (σ, δ). Then $\mathbf{x}_{\delta_{=1}} \coloneqq \tilde{\mathbf{x}}_{\delta_{=1}}$ and $\mathbf{x}_c \coloneqq \boldsymbol{\rho}_r \tilde{\mathbf{x}}_{f(x)}$ for all r such that $c \in \operatorname{supp}(\mathbf{A}_r)$ is well defined and thus $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution and is a feasible solution to the unaggregated problem* (1)*.*

3.1 Sufficient conditions for the optimality of a $\boldsymbol{\rho}$-solution

Until now, we have only considered feasibility, ignoring the relationship between the cost of $\tilde{\mathbf{x}}$ and the cost of $\mathbf{x}$. The following observation gives a condition for the cost of $\tilde{\mathbf{x}}$ to be equal to the cost of $\mathbf{x}$. For all $c \in C \in \delta_{>1}$, let $r_c \in R^{(C)}$ be such that $\mathbf{x}_c = \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_{f(c)}$.

Observation 3.5 *Let $\mathbf{x}$ be a $\boldsymbol{\rho}$-solution. If $\boldsymbol{\omega}_{r_c} \tilde{\mathbf{c}}_{f(c)} = \mathbf{c}_c$ for all $r_c \in R^{(C)} \in \sigma_{>1}$ and $C \in \delta_{>1}$, then the cost of $\tilde{\mathbf{x}}$ for the aggregated problem is equal to the cost of $\mathbf{x}$ in the unaggregated problem.*

Proof Let $\tilde{\mathbf{x}}$ be a solution to the aggregated problem (13). Using Observation 3.3, for all $C \in \delta_{>1}$ the cost corresponding to the variable $\tilde{\mathbf{x}}_C$ is

$$\tilde{\mathbf{c}}_C \tilde{\mathbf{x}}_C = \tilde{\mathbf{c}}_C \left(\sum_{r_c \in R^{(C)}} \boldsymbol{\omega}_{r_c} \boldsymbol{\rho}_{r_c} \right) \tilde{\mathbf{x}}_C = \sum_{r_c \in R^{(C)}} \tilde{\mathbf{c}}_C \boldsymbol{\omega}_{r_c} \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_C = \sum_{c \in C} \mathbf{c}_c \mathbf{x}_c$$

Which corresponds to the cost of the variables C. Thus

$$\tilde{\mathbf{c}}\tilde{\mathbf{x}} = \sum_{C \in \delta_{=1}} \tilde{\mathbf{c}}_C \tilde{\mathbf{x}}_C + \sum_{C \in \delta_{>1}} \tilde{\mathbf{c}}_C \tilde{\mathbf{x}}_C = \sum_{C \in \delta_{=1}} \tilde{\mathbf{c}}_C \tilde{\mathbf{x}}_C + \sum_{C \in \delta_{>1}} \mathbf{c}_C \mathbf{x}_C = \mathbf{c}\mathbf{x}.$$

□

Example 4 (Cont. Example 1) The cost of an aggregated variable of $\text{CEP}_{\mathcal{P}}$ was defined in Subsection 2.2 as equal to the corresponding unaggregated variable, that is $\tilde{\mathbf{c}}_{f(c)} = \mathbf{c}_c$ for all $c \in [n]$. Thus, since the aggregation weights for $\text{CEP}_{\mathcal{P}}$ are all equal to 1, the hypothesis

$$\boldsymbol{\omega}_{r_c} \tilde{\mathbf{c}}_{f(c)} = \mathbf{c}_c$$

of Observation 3.5 trivially holds for $\text{CEP}_{\mathcal{P}}$, for all $r_c \in R^{(C)} \in \sigma_{>1}$ and $C \in \delta_{>1}$.

While row aggregation of a linear problem is a relaxation of the original problem, the same does not apply to column aggregation. In general, it might not even be clear what "relaxation" means since the number of variables changes.

Definition 3.4 Let P and $\tilde{P}$ be a minimization problems of the form $\min_{x\in\mathcal{X}} c(x)$ and $\min_{\tilde{x}\in\tilde{\mathcal{X}}} \tilde{c}(\tilde{x})$. We say that $\tilde{P}$ is a relaxation of P if there exist a function $g : \mathcal{X} \to \tilde{\mathcal{X}}$ such that $\tilde{c}(g(x)) = c(x)$. We refer to g as a cost-preserving function.

With this definition, the column aggregation used for the Capacity Expansion Problem in this work is still a relaxation. In general, a column aggregation of a linear problem is a relaxation of the original problem whenever it is a *constant-coefficients column aggregation*, that is:

Definition 3.5 A column aggregation of a linear problem with respect to the column partition δ is called a *constant-coefficients column aggregation* if, for all $C \in \delta_{>1}$, the following conditions hold:

1. The non-zero rows in $\mathbf{A}^C$ are identical up to multiplication by -1.
2. For all $c \in C$, the cost coefficient satisfies $\mathbf{c}_c = \frac{A^{\mathbf{c}}_r}{\tilde{\mathbf{A}}^C_r}\tilde{\mathbf{c}}_C$, where r is a non-zero row of $\mathbf{A}^C$.

Example 5 (Cont. Example 1) Consider $\text{CEP}_{\mathcal{P}}$ as defined in 2.2. Since the column aggregation consisted of substituting the following summations with an aggregated variable:

$$z^{(e\to h)}_{sIi} \coloneqq \sum_{t\in I} z^{(e\to h)}_{sti}, \qquad z^{(h\to e)}_{sIi} \coloneqq \sum_{t\in I} z^{(h\to e)}_{stn} \tag{27}$$

(and the same is done for the variables z^{Δ}_{sIi}, $z^{(e)}_{sIj}$ and $z^{(h)}_{sIj}$), this is a constant-coefficients columns aggregation.

To see that constant-coefficients column aggregations are relaxations, we construct a cost-preserving function:

Observation 3.6 *Let* $\mathbf{x}$ *be a feasible solution of a linear problem* P*, and consider the aggregated problem* $\tilde{P}$ *obtained by a constant-coefficients column aggregation. Than* $g(\mathbf{x}) \coloneqq \tilde{\mathbf{x}}$*, where* $\tilde{\mathbf{x}}_C \coloneqq \sum_{c\in C} \frac{\mathbf{A}^c_r}{\tilde{\mathbf{A}}^C_r}\mathbf{x}_c$*, where* r *is any non-zero row of* $\mathbf{A}^C$*, is a cost-preserving map and thus* $\tilde{P}$ *is a relaxation of* P*.*

Proof First we check that $g(x)$ is a feasible solution for $\tilde{P}$:

$$\tilde{\mathbf{A}}_r\tilde{\mathbf{x}} = \sum_{C\in\delta} \tilde{\mathbf{A}}^C_r \mathbf{x}_C = \sum_{C\in\delta} \tilde{\mathbf{A}}^C_r \sum_{c\in C} \frac{\mathbf{A}^c_r}{\tilde{\mathbf{A}}^C_r}\mathbf{x}_c = \mathbf{A}_r x = \mathbf{b}_r = \tilde{\mathbf{b}}_r.$$

Lastly the cost of $\tilde{\mathbf{x}} = \tilde{\mathbf{c}}(g(\mathbf{x}))$ is:

$$\sum_{C\in\delta} \tilde{\mathbf{c}}_C\tilde{\mathbf{x}}_C = \sum_{c\in\delta}\sum_{c\in C} \tilde{\mathbf{c}}_C \frac{\mathbf{A}^c_r}{\tilde{\mathbf{A}}^C_r}\mathbf{x}_c = \sum_{c\in\delta}\sum_{c\in C} \mathbf{c}_c\mathbf{x}_c = \mathbf{c}^T\mathbf{x}$$

. □

The optimal value of a relaxation is lower than the optimal value of the original problem, so we have:

Proposition 3.2 *If (σ, δ) is a structure-preserving, constant-coefficients row and column aggregation and $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution such that the hypothesis of Observation 3.5 are satisfied, and the constraints in $\sigma_{=1}$ hold for $\mathbf{x}$, then $\mathbf{x}$ is an optimal solution for the unaggregated problem 1.*

Proof For Observation 3.5, the cost of the aggregated problem is equal to the cost of $\mathbf{x}$ in the unaggregated problem. We only need to show that the aggregated problem is a relaxation of the unaggregated problem. But this is true since both row aggregations and constant-coefficients column aggregations are relaxations. □

We conclude by observing that if all the constraints in $\sigma_{=1}$ satisfy the hypothesis of Observation 3.4, and the cost $\tilde{\mathbf{c}}$ satisfies the hypothesis of Observation 3.5, then every $\boldsymbol{\rho}$-solution is optimal for the unaggregated Problem 1.

3.2 Alternative Iterative definition of $\boldsymbol{\rho}$

A different approach to defining $\boldsymbol{\rho}$ involves selecting an initial aggregated constraint $R_0 \in \mathcal{E}$, assigning $\boldsymbol{\rho}_r$ for each $r \in R_0$ as outlined in Observation 3.2, and setting $\mathbf{x}_c \coloneqq \boldsymbol{\rho}_r \tilde{\mathbf{x}}_{R,f(c)}$ for $c \in C \in \text{supp}(R)$, and $\mathbf{x}_{\delta_{=1}} = \tilde{\mathbf{x}}_{\delta_{=1}}$. Where we denote as $\text{supp}(R) \coloneqq \text{supp}(\tilde{\mathbf{A}}_R)_{>1} \subset \delta_{>1}$, the set of aggregated variables in R, for all $R \in \sigma$. At each step, we then select a new aggregated constraint R and extend the solution $\mathbf{x}$ to the not yet defined variables in $\text{supp}(R)$ as follows. The sets of processed aggregated and unaggregated variables are initialized, respectively, as $J_0 \coloneqq \text{supp}(R)$ and $I_0 \coloneqq \cup_{C \in \text{supp}(R)} C$. We proceed inductively: for $n \geq 1$, $R_n \in \sigma_{>1}$ is chosen such that $\text{supp}(R_n) \setminus J_{n-1} \neq \emptyset$, meaning it contains variables that have not yet been defined, then for each $r \in R_n$, we define:

$$\boldsymbol{\rho}_r \coloneqq \frac{\mathbf{b}_r - \mathbf{A}_{r,I_{n-1}} \mathbf{x}_{I_{n-1}}}{\tilde{\mathbf{b}}_{R_n} - \tilde{\mathbf{A}}_{R_n,J_{n-1}}}$$

and update $I_n \coloneqq \bigcup_{C \in \text{supp}(R_n)} C \cup I_{n-1}$ and $J_n \coloneqq \text{supp}(R_n) \cup J_{n-1}$.

$$\mathbf{x}_c \coloneqq \boldsymbol{\rho}_r \tilde{\mathbf{x}}_{R_n,f(c)}$$

for all $c \in I_n \setminus I_{n-1}$ and $r \in R_n$. The iterations stop at step $N_{max} \in \mathbb{N}$ if $J_{N_{max}} = \delta_{>1}$, that is when the solution $\tilde{\mathbf{x}}$ has been extended to every variable of the unaggregated problem. We refer to a $\boldsymbol{\rho}$ defined by this approach as obtained by *the iterative method* of $\boldsymbol{\rho}_r$. And refer to the corresponding solution $\mathbf{x}$ as a *$\boldsymbol{\rho}$-solution obtained by the iterative method.* Following the same steps as in the proof of Observation 3.2, it can be easily seen that:

Observation 3.7 *Let $\mathbf{x}$ be a $\boldsymbol{\rho}$-solution defined by the iterative method. Then $\mathbf{x}$ is feasible for all constraints $r \in R_n$ for $n = 1, \ldots, N_{max}$.*

We observe that the advantage of this iterative definition of $\boldsymbol{\rho}$, over the method described in Observation 3.2 is that a $\boldsymbol{\rho}$-solution obtained by the iterative method $\mathbf{x}$, is always well defined whenever the following assumption holds:

Assumption 2 *For every $r, r' \in R \in \sigma_{>1}$, we have*

$$\mathrm{supp}(\mathbf{A}_r)_{>1} \cap \mathrm{supp}(\mathbf{A}_{r'})_{>1} = \emptyset.$$

Assumption 2 is weaker than requiring $\boldsymbol{\rho}_r$ to be constant on the connected components of the Hypergraph associated to the partition (σ, δ) as in Proposition 3.1. From now on, we will assume that assumption 2 always holds. We now extend Observation 3.3. Instead of Assumption 1, we need the following Assumption:

Assumption 3 *If $C \in \mathrm{supp}(R)$, then $C \subset \cup_{r \in R}(\mathrm{supp}(\mathbf{A}_r))$.*

Assumption 3 ensures that if an aggregated variable C is in the support of $\tilde{\mathbf{A}}_R$, then every corresponding unaggregated variable $c \in C$, appears in at least one constraint $r_c \in R$. In the following observation, we use the sequence $R_0, \ldots, R_{N_{max}}$ as defined in the iterative definition of $\boldsymbol{\rho}$.

Observation 3.8 *If $\boldsymbol{\rho}$ is defined by the iterative method, if for $n = 0, \ldots, N_{max}$, for all $i_k < n$ such that $\mathrm{supp}(R_{i_k}) \cap \mathrm{supp}(R_n) \neq \emptyset$ we have $\boldsymbol{\omega}_{R_n} = \boldsymbol{\omega}_{R_{i_k}}$, then for all R_n we have:*

$$\boldsymbol{\rho}_{R_n} \boldsymbol{\omega}_{R_n} = 1 \tag{28}$$

Proof We proceed by induction on n. This is true for $n = 0$ because of Observation 3.7. For $n > 0$, we have:

$$\sum_{r \in R_n} \boldsymbol{\omega}_r \boldsymbol{\rho}_r = \frac{\sum_{r \in R} \boldsymbol{\omega}_r \mathbf{b}_r - \sum_{r \in R_n} \boldsymbol{\omega}_r \mathbf{A}_{r,\delta=1} \mathbf{x}_{\delta=1} - \sum_{r \in R_n} \mathbf{A}_{r,I_{n-1}} \mathbf{x}_{I_{n-1}}}{\tilde{\mathbf{b}}_{R_n} - \tilde{\mathbf{A}}_{R_n,\delta=1} \tilde{\mathbf{x}}_{\delta=1} - \tilde{\mathbf{A}}_{R_n,J_{n-1}} \tilde{\mathbf{x}}_{J_{n-1}}} \tag{29}$$

So, we want to show that the numerator is equal to the denominator. Consider the sum:

$$\sum_{r \in R} \mathbf{b}_r - \sum_{r \in R_n} \boldsymbol{\omega}_r \mathbf{A}_{r,\delta=1} \mathbf{x}_{\delta=1} - \sum_{r \in R_n} \mathbf{A}_{r,I_{n-1}} \mathbf{x}_{I_{n-1}} \tag{30}$$

The fist and the second sum are equal to $\tilde{\mathbf{b}}_{R_n}$ and $\tilde{\mathbf{A}}_{R_n} \tilde{\mathbf{x}}_{\delta=1}$ respectively. Then, since $J_{n-1} = \cup_{i=1}^{n-1} \cup_{C \in \mathrm{supp}(R_i)} \{C\}$, for all $C \in J_{n-1}$, there exists some $i_C \leq n-1$ such that $\mathbf{x}_c = \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_{f(c)}$ for all $r_c \in R_{i_C}$. The last sum is equal to:

$$\sum_{r \in R_n} \boldsymbol{\omega}_r \mathbf{A}_{r,I_{n-1}} \mathbf{x}_{I_{n-1}} = \sum_{r \in R_n} \sum_{C \in J_{n-1}} \boldsymbol{\omega}_r \mathbf{A}_{r,C} \mathbf{x}_C = \sum_{r \in R_n} \sum_{C \in J_{n-1}} \boldsymbol{\omega}_r \mathbf{A}_{r,C} \boldsymbol{\rho}_{R_{i_C}} \tilde{\mathbf{x}}_C \tag{31}$$

Since only one coefficient in $\mathbf{A}_{r,C}$ is not null, and $\boldsymbol{\omega}_{R_n} = \boldsymbol{\omega}_{i_C}$ by hypothesis, we have:

$$= \sum_{C \in J_{n-1}} \tilde{\mathbf{A}}_{R,C} \tilde{\mathbf{x}}_C \sum_{r \in R_{i_C}} \boldsymbol{\omega}_r \boldsymbol{\rho}_r = \sum_{C \in J_{n-1}} \tilde{\mathbf{A}}_{R_n,C} \tilde{\mathbf{x}}_C \tag{32}$$

Where the last equality comes from the hypothesis, $\boldsymbol{\omega}_{R_n} = \boldsymbol{\omega}_{R_{i_C}}$ if $\tilde{\mathbf{A}}_{R_n,C} \neq 0$ (Since $\tilde{\mathbf{A}}_{R_n,C} \neq 0$ implies that $\text{supp}(R_C) \cap \text{supp}(R_n) \neq \emptyset$) . So we have:

$$\sum_{r \in R_{i_C}} \boldsymbol{\omega}_r \boldsymbol{\rho}_r = \boldsymbol{\omega}_{R_{i_C}} \boldsymbol{\rho}_{R_{i_C}}$$

which is equal to 1 by induction. □

If $\mathbf{x}$ is a $\boldsymbol{\rho}$ solution defined by the iterative method, then for all $C \in \delta_{>1}$ there exists $i_C \leq n_{\max}$ such that $\mathbf{x}_c \coloneqq \boldsymbol{\rho}_{r_c} \tilde{f(c)}$ for all $r_c \in R_{i_C}$. Thus we can easily see that Observation 3.4 and Observation 3.5 become respectively:

Observation 3.9 *Let (σ, δ) be a structure-preserving aggregation, such that the hypothesis from Observation 3.8 holds. If $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution defined by the iterative method and r is a constraint in $\sigma_{=1}$, such that*

$$\mathbf{A}_{r,c} = \boldsymbol{\omega}_{r_c} \tilde{\mathbf{A}}_{r,f(c)} \text{ for all } C \in \delta_{>1},\ r_c \in R_{i_C},$$

then $\mathbf{x}$ is a feasible solution for constraint r.

Observation 3.10 *Let (σ, δ) be a structure-preserving aggregation, such that the hypothesis from Observation 3.8 holds. Let $\mathbf{x}$ be a $\boldsymbol{\rho}$-solution If $\boldsymbol{\omega}_{r_c} \tilde{\mathbf{c}}_{f(c)} = c_c$ for all $r_c \in R_{i_C}$ for all $C \in \delta_{>1}$, then the cost of $\tilde{\mathbf{x}}$ for the aggregated problem is equal to the cost of $\mathbf{x}$ in the unaggregated problem.*

From this follows the following sufficient condition for a $\boldsymbol{\rho}$-solution $\mathbf{x}$ obtained by the iterative method to be optimal for the unaggregated problem:

Proposition 3.3 *Let (σ, δ) be a structure-preserving, constant-coefficients aggregation, such that the hypothesis from Observations 3.8 and 3.10 hold. If $\mathbf{x}$ is a $\boldsymbol{\rho}$-solution defined by the iterative method and is feasible for the constraints $r \in R$ with $R \neq R_n$ for all n, then $\mathbf{x}$ is optimal for the unaggregated problem.*

We conclude with a comment on the applicability of Proposition 3.3. The set of constraints r in R with $R \neq R_n$ for all n corresponds to the set of unaggregated constraints and the set of aggregated constraints different from all R_n. If the former set is comprised of constraints of the type described in Observation 3.9, then these are always satisfied for a $\boldsymbol{\rho}$-solution defined by the iterative method. Furthermore, if every constraint has an aggregated variable that is not present in any other constraint, then the set $\{R \in \sigma_{>1} \mid R \neq R_n \text{ for } n = 1, \ldots, N_{max}\}$ is empty. Finally, we note that all results in this section remain valid even if some or all of the unaggregated variables are constrained to integer values.

4 Aggregated Capacity Expansion Problem

In this section, we apply the results of Section 3 to the time series aggregation model $\text{CEP}_{\mathcal{P}}$.

As discussed in Example 1, $\text{CEP}_{\mathcal{P}}$ is a *structure-preserving* row and column aggregation (see Definition 3.1) of $\text{CEP}_{\mathcal{T}}$ without the intermediate Hydrogen Storage Constraints (5). The function f maps variables with a time index $t \in \mathcal{T}$ to the aggregated variables with a time index $I \in \mathcal{P}$ where $t \in I$. For instance, for the time partition $\mathcal{P} \coloneqq \{[1,2],[3,4,5],[6]\}$, we have:

$$f(z_{1is}^{(e\to h)}) = z_{[1,2]is}^{(e\to h)}, \quad f(z_{2is}^{(e\to h)}) = z_{[1,2]is}^{(e\to h)},$$

and $f(x^{(p)}) = x^{(p)}$, and analogously, the same is done for all the other variables. Furthermore, as discussed in Example 5, $\text{CEP}_{\mathcal{P}}$ is a constant-coefficients columns aggregation (see Definition 3.5). Thus, for Observation 3.6, the aggregated problem $\text{CEP}_{\mathcal{P}}$ is a relaxation of $\text{CEP}_{\mathcal{T}}$.

The above discussion generally holds in the case of any pair of time partitions $\mathcal{P}$ and $\mathcal{P}'$ where $\mathcal{P}'$ is finer than $\mathcal{P}$. We can summarize the above in the following observation.

Proposition 4.1 *Let $\mathcal{P}$ and $\mathcal{P}'$ be two time partitions of $\mathcal{T} = \{1,\dots,T\}$ such that $\mathcal{P}'$ is finer than $\mathcal{P}$. Then, $CEP_{\mathcal{P}}$ is a relaxation of $CEP_{\mathcal{P}'}$, and the optimal solution to $CEP_{\mathcal{P}}$ provides a lower bound for the optimum to $CEP_{\mathcal{P}'}$.*

As shown in Example 3, Assumption 1 holds for $\text{CEP}_{\mathcal{P}}$. Furthermore, since the weights of the row aggregation are all equal to 1, the hypothesis $\boldsymbol{\omega}_{r_c}\tilde{\mathbf{c}}_{f(c)} = \tilde{\mathbf{c}}_{f(c)} = c_c$ in Observation 3.5 holds. First, we observe, as seen in Section 3.1, that a $\boldsymbol{\rho}$-solution is well defined if and only if $\boldsymbol{\rho}_r$ is constant on the connected components of the hypergraph associated with the aggregation. Since $\text{CEP}_{\mathcal{P}}$ has exactly one connected component for each time-step t in $\mathcal{T}$, we refer to the value of $\boldsymbol{\rho}_r$ on the connected component corresponding to t, as $\boldsymbol{\rho}_t$.

Lastly, for Proposition 3.2, a $\boldsymbol{\rho}$-solution $\mathbf{x}$ is optimal for $CEP_{\mathcal{T}}$ if the unaggregated constraints $r \in \sigma_{=1}$ hold. Unaggregated constraints of $\text{CEP}_{\mathcal{P}}$ always hold for $\boldsymbol{\rho}$-solutions: this is trivial for unaggregated Electricity Balance (3) and Hydrogen Balance (4') constraints, since no aggregated variable is in their support. As seen in Example 2, the periodic Hydrogen Storage Constraint (6) holds for a $\boldsymbol{\rho}$-solution. Moreover, since we initially removed intermediate Hydrogen Storage Constraints (7) to obtain $\text{CEP}_{\mathcal{P}}$, we now check that a $\boldsymbol{\rho}$-solution also holds for these constraints. Let us fix a time interval $I \coloneqq I_{h'} = \{t_1,\dots,t_{|I|}\} \in \mathcal{P}$ and an intermediate time-step $t_{\bar{k}} \in I_{h'}$. We assume that the aggregated variable $\tilde{z}_{t_k}^{\Delta}$ is greater than 0, but a similar argument also holds if $\tilde{z}_{t_k}^{\Delta} \leq 0$. The intermediate storage value of the $\boldsymbol{\rho}$-solution at time $t_{\bar{k}}$ is equal

to:

$$z_{t_{\bar{k}}} \coloneqq z_{t_1-1} + \sum_{k=1}^{\bar{k}-1} z^{\Delta}_{t_k} \tag{33}$$

$$= \tilde{z}_{I_{\bar{h}-1}} + \sum_{k=1}^{\bar{k}-1} \boldsymbol{\rho}_{t_k} \tilde{z}^{\Delta}_{I_{h'}} \tag{34}$$

$$= \tilde{z}_{I_{\bar{h}-1}} + \sum_{k=1}^{|I_{h'}|} \boldsymbol{\rho}_{t_k} \tilde{z}^{\Delta}_{I_{h'}} \tag{35}$$

$$= \tilde{z}_{I_{\bar{h}-1}} + \tilde{z}^{\Delta}_{I_{h'}} \tag{36}$$

$$\leq x^{(h)}, \tag{37}$$

where (34) comes from the definition of $\boldsymbol{\rho}$-solution, (35) holds since $\boldsymbol{\rho} \geq 0$ and $\tilde{z}^{\Delta}_{t_k} \geq 0$ and (36) holds since $\sum_{k=1}^{|I_{h'}|} \boldsymbol{\rho}_{t_k} = 1$ for Observation 3.3. Finally inequality (37), corresponds to the Hydrogen Storage Constraint (7) for time-step $I_{h'}$, for the aggregated problem $\text{CEP}_{\mathcal{P}}$. Note that the intermediate Hydrogen Storage Constraints also hold for $\boldsymbol{\rho}$-solutions. Therefore, we have the following result.

Proposition 4.2 *A $\boldsymbol{\rho}$-solution $\mathbf{x}$ is an optimal solution for $CEP_{\mathcal{T}}$.*

The same arguments can be applied to a $\boldsymbol{\rho}$-solution obtained with the iteration method.

Proposition 4.3 *A $\boldsymbol{\rho}$-solution obtained by the iteration method $\mathbf{x}$, is an optimal solution for $CEP_{\mathcal{T}}$.*

The propositions above establish that $\boldsymbol{\rho}$-solutions are not only feasible but also optimal for $\text{CEP}_{\mathcal{T}}$. In the following section, we leverage this insight to develop an iterative approach to refining time partitions, progressively converging to an optimal solution for the original problem.

5 Iterations on time partitions

Since any time-aggregated LP formulation provides a relaxation to the original LP, but with significantly fewer variables and reduced optimization time, we aim to employ carefully selected aggregations to iteratively warm-start the solver, progressively converging to the optimal solution through increasingly refined aggregations.

Power generation and electricity load data typically exhibit very strong seasonal and daily patterns and, to a minor extent, a weekly trend. Generally, long-term patterns can be captured even with relatively coarse aggregations. However, a grid described by the solution to the $\text{CEP}_{\mathcal{P}}$for a loose partition will likely be unable to deal with daily variability. For instance, it is common to experience days with overall greater power production than load, but with

peak production occurring at noon and most of the power load during the late evening, the aggregated model overlooks such a discrepancy.

Hence, the method devised is the following:

1. Set up the model environment with enough variables for future iterations. Impose the constraints relative to an initial time partition and solve the corresponding LP.
2. Select a time interval using a specified selection criteria and refine it into smaller sub-intervals.
3. Add the constraints relative to each sub-interval of the selected interval. Solve the model again but using a warm-start.
4. Repeat steps 2 and 3 until a specified halting condition is met.

Our implementation refines each interval to 1-hour time-steps and we consider all scenarios to share the same time partition. However, exploring alternative refinement methods, such as splitting intervals into larger than 1 hour subintervals, and reducing the number of constraints introduced at each iteration by considering each scenario with a different time partition, are priorities for future work. As a baseline, we have implemented an algorithm that uses random selection criteria and stops after a fixed number of iterations.

5.1 Iterations based on $\boldsymbol{\rho}$

We now present a second method, based on the index $\boldsymbol{\rho}$ as defined in Section 3, for selecting the day to disaggregate in step 2 of the algorithm described in Section 5. Proposition 4.2 states that a $\boldsymbol{\rho}$-solution $\mathbf{x}$ is optimal for $\mathrm{CEP}_{\mathcal{T}}$. However, a $\boldsymbol{\rho}$-solution is well defined if and only if the following two conditions hold for $\boldsymbol{\rho}$:

1. $\boldsymbol{\rho}_r$ is constant over the hypergraph associated to the aggregated problem $\mathrm{CEP}_{\mathcal{P}}$.
2. $\boldsymbol{\rho} \geq 0$.

Where Condition (2) corresponds to the requirement that magnitude constraints hold for the $\boldsymbol{\rho}$-solution. This can be easily seen in Linear Problem 1, where if $\boldsymbol{\rho}_{r_c} < 0$, then $\mathbf{x}_c \coloneqq \boldsymbol{\rho}_{r_c} \tilde{\mathbf{x}}_C < 0$ would violate the positivity constraint. Consider the constraints r^{P}_{ist} and r^{H}_{ist} corresponding to the Electricity Balance (3) and Hydrogen Balance (4), for node i, scenario s at time $t \in I \in \mathcal{P}$ with $|I| > 1$. From the definition of $\boldsymbol{\rho}$, we can see that $\boldsymbol{\rho}_{r^{\mathrm{P}}_{ist}}$ is equal to the ratio between the net power production at the node at time t, divided by the total net power production at node i during the whole time interval I. Then, Condition 1 corresponds to $\boldsymbol{\rho}_{r^{\mathrm{H}}_{ist}} = \boldsymbol{\rho}_{r^{\mathrm{P}}_{ist}}$ since the variables ${z^{(e\to h)}}_{ist}$ appear in both constraints and also that $\boldsymbol{\rho}_{r^{\mathrm{H}}_{ist}} = \boldsymbol{\rho}_{r^{\mathrm{H}}_{i',st}}$ for all nodes $i' \in \mathcal{N}$ (assuming $\mathcal{G}$ is a connected graph) because of the binding variables between adjacent nodes ${z^{(e)}}_{lst}$. That is, for a $\boldsymbol{\rho}$-solution to be well defined, fixed a scenario s, and a time-step t, the ratio of the net energy production should be equal for all nodes in the network and equal to the ratio of the net hydrogen production.

Since constraints across different scenarios do not share aggregated variables, each scenario can have its own distinct time partition. This approach reduces the number of constraints added at each iteration.

We observe that these conditions are no longer required if we instead use $\boldsymbol{\rho}$ as defined by the iterative method in Subsection 3.2.

However, while $\boldsymbol{\rho}_r$ initially corresponded to fractional net power production with clear physical significance, the iterative method definition of $\boldsymbol{\rho}_r$ lacks this interpretability.

These conditions provide the following heuristic selection criteria for choosing the intervals to refine, with $\boldsymbol{\rho}$ defined as in Section 3:

1. Refine $I \in P$ where $\boldsymbol{\rho}_r$ shows the greatest variance across nodes constraining the same time-step $t \in I$ and scenario s.
2. If $\boldsymbol{\rho}_r$ changes sign over the time interval I, refine I into smaller intervals so that $\boldsymbol{\rho}_r$ maintains the same sign within each sub-interval.

In contrast, if we used $\boldsymbol{\rho}$ defined by the iterative method, only Heuristic 2 would be applicable, as $\boldsymbol{\rho}$-solution defined by the iterative method is well-defined even if $\boldsymbol{\rho}$ has different values across the nodes for a fixed time-step.

Lastly, if the heuristics do not identify any interval I meeting the conditions, then for Propositions 4.2 and 4.3, we have a $\boldsymbol{\rho}$-solution that is optimal for the original problem $\text{CEP}_{\mathcal{T}}$.

5.2 Iterations based on Rolling Horizon

While computing the optimal solution on a batch of scenarios by solving the LP model described in Section 2, the solver has perfect foresight. Specifically, the solver determines the optimal electricity and hydrogen conversion or transmission levels at each time-step using a mathematical minimum that incorporates the entire year's demand data. In practice, accurate forecasts for weather, and consequently for power generation, are known on a day-ahead basis, at most two days ahead. Thus, we aim to assess whether an optimal grid configuration, as derived from the solution of the CEP, can function with limited foresight to meet demand on the scenarios $\mathcal{J}$. On a single-node micro-grid, a deterministic system control can be designed and used to assess whether a single-node CEP solution is sufficient for feasibility over a given scenario, even without day-ahead forecasts. Such a strategy is, for example, discussed in more detail in Wang and Nehrir (2008). However, this is not extendable to the case of a multiple-node network structure. In order to operate a multi-node grid, we propose to use the Rolling Horizon optimization technique. This technique divides the time horizon into smaller periods and sequentially optimizes each period, treating the solved periods as fixed for the next. In our case, each period spans one day, corresponding to the time-frame provided by daily weather forecasts.

Consider a solution $\mathbf{x}$ of $\text{CEP}_{\mathcal{P}}$, and consider a test scenario $s \in \mathcal{J}$ with its generation and load time series. The Rolling Horizon algorithm is outlined as follows:

1. **Initialization:** Set ${z_{0i}}^{test}$ equal to the maximum initial storage z_{s0n} over all scenarios s in $\mathcal{J}$.
2. **Daily Iteration:** For each day d in the time horizon:
   - Optimize the inner Economic Dispatch problem for the given day, with initial hydrogen storage values equal to the initial storage of the solution $\mathbf{x}$ of $\text{CEP}_{\mathcal{P}}$, that is ${z_{d-1i}}^{test}$. If the problem is infeasible, terminate the process.
   - Update the hydrogen storage levels ${z_{in}}^{test}$ equal to the storage at the last time-step of the day of the current solution of the Economic Dispatch Problem.

The daily ED problem is defined analogously to the CEP problem described in Section 2, with the difference that the first-stage variables (number of generators, carrying capacities of the power and hydrogen lines, and power cells capacity) are fixed, and the hydrogen storage level does not loop as in the full year LP but is given an initial hydrogen storage from a previous computation.

Definition 5.1 (RH-feasibility) Given a solution $\mathbf{x}$ to the Capacity Expansion Problem and given a test scenario $\hat{s}$, we consider $\mathbf{x}$ to be RH-feasible over scenario $\hat{s}$ if the Rolling Horizon optimization algorithm terminates at the end of the year and ${z_{iT}}^{test} \geq {z_{i0}}^{test}$ for all nodes $i \in \mathcal{N}$.

We include the final condition to ensure that the hydrogen storage levels at the end of the year are at least as high as those at the beginning. This requirement flags as infeasible any solutions that satisfy demand throughout the year solely by relying on a net consumption of unproduced hydrogen.

When formulated in this manner, the daily optimization discourages the storage of hydrogen unless it is required within the same day due to the associated costs of operating the electrolyzers. This tendency can lead to the infeasibility of scenarios that would otherwise be feasible with more effective storage management. To address this issue, we introduce a discrepancy function into the model. In our solution, we assign a cost to the difference between the hydrogen storage level at time-step t and the average of the corresponding variables from the optimal solutions of the training scenarios. Thus we define positive variables discrepancy_{it} for each time-step t and node i, with positive cost, and add the constraints:

$$\text{discrepancy}_{it} \;\geq\; \frac{1}{J}\left(\sum_{s\in\mathcal{J}} z_{ist}\right) - {z_{it}}^{test}. \tag{38}$$

Alternatively, one could assign a slight negative cost to ${z_{it}}^{test}$ to incentivize the filling of the storage. However, this approach risks inflating the estimated total cost by operating electrolyzers more than necessary.

We observe that the overall solution to the Economic Dispatch problem over the full-time horizon, obtained using the RH method, is not necessarily optimal, whether or not the additional discrepancy function is included. While

the feasibility of the RH solutions ensures that the first-stage solutions derived from a solution $\tilde{\mathbf{x}}$ of the aggregated problem $\text{CEP}_{\mathcal{P}}$ are feasible, the complete solution—considering second-stage variables associated with hydrogen conversion, reconversion, and transportation—may not be feasible for $\text{CEP}_{\mathcal{T}}$. Nevertheless, some insights regarding this solution's distance from the optimal ED solution can be derived. Specifically, based on the findings of Glomb et al. (2022), we know that a bound can be derived on the ratio between the perfect foresight optimum and RH solution, dependent on $x_i^{(h\to e)}$. To achieve a solution to the ED problem that is closer to the perfect foresight optimum, more refined RH techniques can be employed, such as optimizing on two-day forecasts with a daily refresh rate.

The disaggregation method we consider utilizes the Rolling Horizon as follows: at each iteration, we validate the first-stage $\mathbf{x}$ variables obtained from the previous iteration by checking for RH-feasibility over the entire year at an hourly time resolution, beginning with one of the training scenarios. If the RH algorithm fails before reaching the end of the year, we designate the day on which the failure occurred as the day to disaggregate. Subsequently, we solve the $\text{CEP}_{\mathcal{P}}$ problem using the new aggregation and restart the validation process with the RH, beginning from the day of the previous failure. If the RH successfully navigates through the entire year for all scenarios, the considered $\mathbf{x}$ values will form part of a feasible solution for the $\text{CEP}_{\mathcal{T}}$ problem over the complete horizon at the finest time resolution.

An advantage of using the RH validation method within the iterative aggregation method is that it automatically provides an effective halting method that guarantees feasibility for the original problem. A drawback of using the RH validation within the iteration loop is its considerable computational cost.

6 Computational Results

To evaluate the methodologies presented in this paper, we examine a 5-node network over a one-year span, utilizing time-steps of 1 hour across two distinct scenarios. The scenarios are generated as described in Appendix A, with a detailed explanation of parameter selection provided in Appendix B. The computational tests were conducted on an Intel(R) Core(TM) i7-13700H CPU @ 2.40GHz with 16 GB of RAM using Gurobi (Gurobi Optimization, LLC 2024).

We compare the three approaches for iterating on the aggregated problem: (1) randomly selecting the interval for refinement, (2) selecting the interval with the highest $\boldsymbol{\rho}$-variance as defined in Subsection 5.1, and (3) selecting the interval where the RH validation method discussed in Subsection 5.2 fails.

Figure 2 illustrates the cost variation at each iteration using the $\boldsymbol{\rho}$ selection method compared to random interval selection. The $\boldsymbol{\rho}$ selection method demonstrates a slightly faster increase in cost than the random selection method, with comparable optimization times: 174 seconds for the $\boldsymbol{\rho}$ selection method and 155 seconds for the random selection method over 10 iterations.

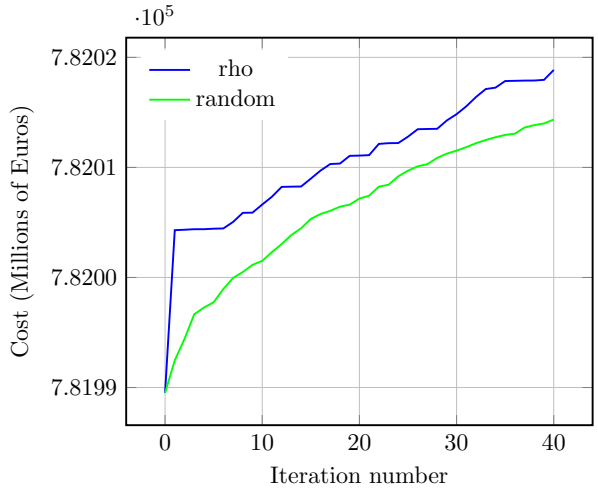

Fig. 2 Cost over iterations of rho selection method versus random selection method

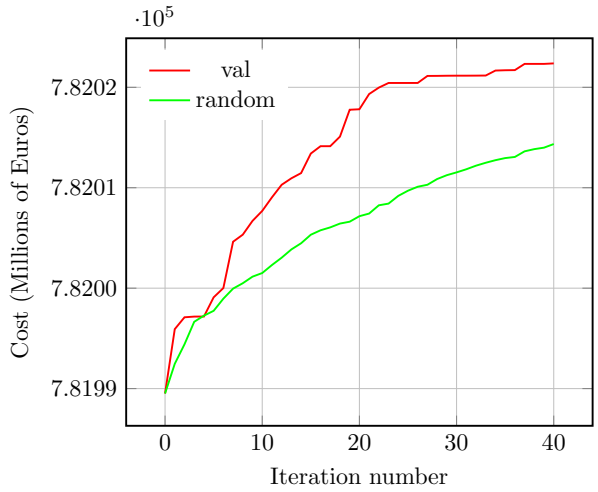

Fig. 3 Cost over iterations of random selection versus rolling horizon selection

In Figure 3, we compare the cost variation at each iteration using the RH validation method for interval refinement against the random interval selection method. The results indicate that the former method yields a significantly faster increase, especially in the first 20 iterations, implying quicker convergence to the optimal solution. However, this approach incurs greater computational time: the RH method takes an average of 137 seconds to complete one full year for the tested scenarios, and iterations can take more than a minute each.

It is worth noting that while both the $\boldsymbol{\rho}$ and random iteration methods continue up until reaching the maximum iteration limit, the RH validation

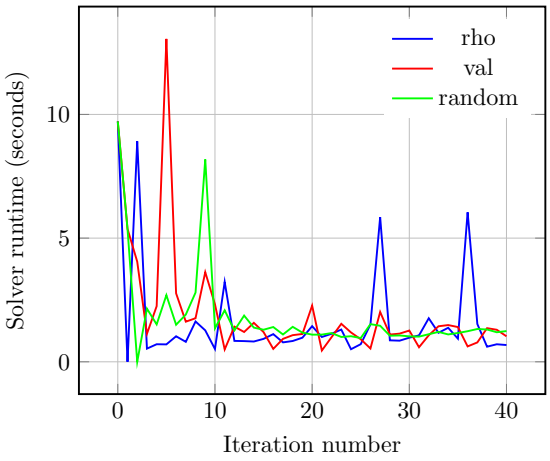

Fig. 4 Solver runtime time over iterations

method iteration may halt before that if the current solution is RH-feasible on the full disaggregated horizon. The solution found is not necessarily optimal, and other heuristics would then be needed to improve it by readjusting the operational costs of hydrogen conversion and transportation. Such adjustments can be made without further increasing the number of disaggregated days, keeping the problem size limited.

However, this advantage does not consistently hold in practice. In many cases, disaggregating the interval on which the RH validation failed did not guarantee that the next iteration would succeed on that interval. In such cases, a random disaggregation would occur for the next iteration. As a result, after the first few efficient iterations, the method often performed similarly to a (very costly) random selection process without reliably improving feasibility.

Figure 4 shows the optimization time across iterations. We observe a generally constant yet slightly decreasing trend in optimization time, indicating that the model is effectively exploiting the warm start. Furthermore, the reduction in optimization time suggests that the number of pivots needed to recover the optimal solution decreases as the optimization progresses, bringing the solution closer to optimality with each iteration. While the optimization time remains similar across the different iteration methods, in Figure 5, we observe significant variation in the total iteration time. This includes the time spent selecting the time interval to disaggregate, adding additional constraints, and reoptimizing. The validation method performs poorly in this regard due to the time consumed by the RH.

Most Capacity Expansion solutions increase only marginally throughout the iterations for all methods considered. Variables $x_i^{(p)}$, $x_i^{(w)}$, and $x_i^{(h)}$ tend to be solved successfully by the initial optimization without the need for any disaggregation. This result is consistent with expectations since the aggregated problem adequately represents overall generation needs and seasonal trends,

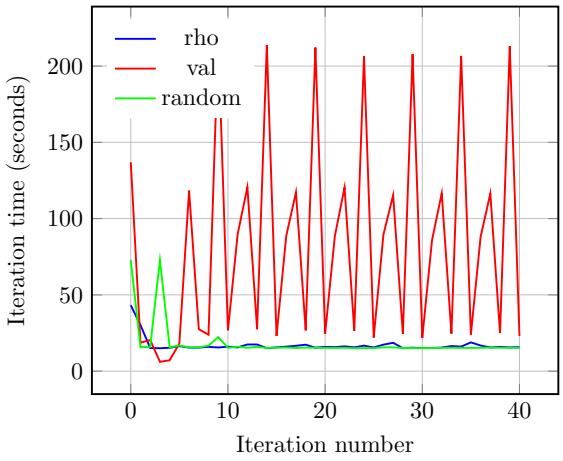

Fig. 5 Iteration time over iterations

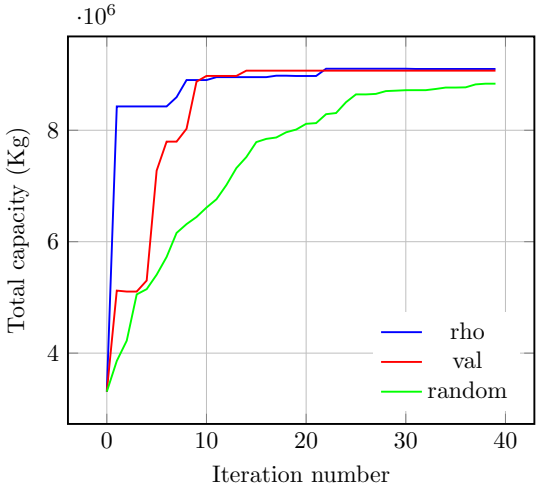

Fig. 6 Hydrogen to Power Capacity over iterations

which dictate the maximum storage needed. On the other hand, variables $x_i^{(h\to e)}$, $x_i^{(e\to h)}$, and $y^{(e)}$ see the greatest increase throughout iterations. In particular, Figure 6 shows the total required hydrogen-to-electricity conversion capacity needed over all nodes, according to the solutions obtained with the three methods. We observe that the $\boldsymbol{\rho}$ index is very efficient at determining a sample day whose disaggregation results in a great improvement in the estimate of the $x_i^{(h\to e)}$ variable. This is to be expected since we specifically devised the index to target high intra-day variability.

7 Conclusion and Future Directions

The computational results presented in the previous sections demonstrate that aggregating time-steps, combined with iterative refinement, can effectively solve the Capacity Expansion Problem (CEP).

We have explored two different approaches for selecting the time intervals to refine at each iteration. The first approach employed a validation method based on the Rolling Horizon approach, which offers the advantage of providing a feasibility certificate if it halts before reaching the maximum number of iterations and reflects a realistic setting where reliable forecasts for power production are available over a short time span.

The second approach utilized the parameter $\boldsymbol{\rho}$ (see Observation 3.2), representing the fraction of net power production in each time-step relative to the total net power production within the corresponding interval. The variance of $\boldsymbol{\rho}$ across each node in the network served as a quality index for each time interval, guiding the disaggregation process. We also provided a theoretical justification for using $\boldsymbol{\rho}$, explaining that time intervals with greater oscillations in net power production require a finer time partition to be accurately considered and establishing sufficient conditions under which the aggregated solution can be extended to an optimal solution for the original problem.

The iteration method with RH validation proved to be initially very effective at selecting relevant intervals to disaggregate, but the quality of the selection diminished quickly with iterations. Furthermore, the use of RH within the iteration loop has a non negligible computational cost, comparable to the cost of solving the full disaggregated LP directly. Yet there is room for improvement within the implementation, especially if considered coupled with other methods that would allow it to overcome those intervals on which it fails repeatedly.

The iteration method with $\boldsymbol{\rho}$ proved to be more effective than chance at selecting relevant intervals to disaggregate while maintaining similar runtimes. In particular, the index was able to efficiently identify time-steps whose disaggregation brought to an accurate estimation of the needed hydrogen-to-electricity conversion capacity in fuel cells. All variables in this study are treated as continuous for ease of testing; however, the iterative method and the optimality results remain valid even when the unaggregated variables are restricted to integer values. In future work, we plan to relax the conditions in Proposition 3.2 for the feasibility of the aggregated solution for the original problem. Furthermore, the interval selection method can be refined by considering other indices based on $\boldsymbol{\rho}$, such as the frequency of sign changes over time within a time interval or by using other measures instead of the variance across the nodes of the network. Lastly, alternative refinement methods can be explored, such as splitting intervals into larger than 1 hour subintervals, and reducing the number of constraints introduced at each iteration by considering each scenario with a different time partition.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

No funding was received for this study.

Availability of data and materials

The dataset containing renewable energy outputs for European countries is available at:
`https://www.renewables.ninja/`.
The dataset containing electricity load data is taken from the ENTSOE platform and is available at:
`https://www.entsoe.eu/data/power-stats/`.
Additional data supporting the findings of this study are provided within the manuscript and supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

G.R. and B.U. wrote the main manuscript text. G.R. developed the theoretical framework, implemented the iterative refinement of the time aggregation, and prepared all the figures. B.U. developed the Rolling Horizon approach, contributed significant insights into the theoretical development, wrote the Parameters and Model Description section, and implemented the Rolling Horizon validation method. S.G. was responsible for designing the notation used throughout the manuscript and carried out in-depth reviews and corrections. All authors read, reviewed, and approved the final version of the manuscript.

Acknowledgements Gabor Riccardi and Stefano Gualandi acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title HEXAGON: Highly–specialized EXact Algorithms for Grid Operations at the National level – CUP F53D23010010001 - Grant Assignment Decree No. 1379 adopted on 01/09/2023 by the Italian Ministry of Ministry of University and Research (MUR).

References

Biener W, Garcia Rosas KR (2020) Grid reduction for energy system analysis. Electric Power Systems Research 185:106349, DOI https://doi.org/10.1016/j.epsr.2020.106349

Blanco H, Faaij A (2018) A review at the role of storage in energy systems with a focus on power to gas and long-term storage. Renewable and Sustainable Energy Reviews 81:1049–1086, DOI https://doi.org/10.1016/j.rser.2017.07.062

Dawood F, Anda M, Shafiullah G (2020) Hydrogen production for energy: An overview. International Journal of Hydrogen Energy 45(7):3847–3869, DOI https://doi.org/10.1016/j.ijhydene.2019.12.059

Domínguez-Muñoz F, Cejudo-López JM, Carrillo-Andrés A, Gallardo-Salazar M (2011) Selection of typical demand days for chp optimization. Energy and Buildings 43(11):3036–3043, DOI https://doi.org/10.1016/j.enbuild.2011.07.024

Entso-e Power Statistics and Transparency Platform (2024) Entso-e power statistics and transparency platform - cross-border physical flow. URL `https://transparency.entsoe.eu/transmission-domain/physicalFlow/show?name=&defaultValue=false&viewType=TABLE&areaType=BORDER_CTY&atch=false&dateTime.dateTime=22.07.2024+00:00|CET|DAY&border.values=CTY|10YGR-HTSO-----Y!CTY_CTY|10YGR-HTSO-----Y_CTY_CTY|10YIT-GRTN-----B&dateTime.timezone=CET_CEST&dateTime.timezone_input=CET+(UTC+1)+/+CEST+(UTC+2)`

ENTSO-E Statistical Reports (2024) Entso-e power statistics and transparency platform. URL `https://www.entsoe.eu/data/power-stats/`

European Commission B (2024) Guidance on article 20a on sector integration of renewable electricity of directive (eu) 2018/2001 on the promotion of energy from renewable sources, as amended by directive (eu) 2023/2413 [c(2024) 5041 final]. URL `https://energy.ec.europa.eu/document/download/efcd200c-b9ae-4a9c-98ab-73b2fd281fcc_en?filename=C_2024_5041_1_EN_ACT_part1_v10.pdf`

European Hydrogen Observatory (2023) URL `https://observatory.clean-hydrogen.europa.eu/tools-reports/levelised-cost-hydrogen-calculator`

Fazlollahi S, Bungener SL, Mandel P, Becker G, Maréchal F (2014) Multi-objectives, multi-period optimization of district energy systems: I. selection of typical operating periods. Computers and Chemical Engineering 65:54–66, DOI https://doi.org/10.1016/j.compchemeng.2014.03.005

Glomb L, Liers F, Rösel F (2022) A rolling-horizon approach for multi-period optimization. European Journal of Operational Research 300(1):189–206, DOI https://doi.org/10.1016/j.ejor.2021.07.043

Gurobi Optimization, LLC (2024) Gurobi Optimizer Reference Manual. URL `https://www.gurobi.com`

Hörsch J, Brown T (2017) The role of spatial scale in joint optimisations of generation and transmission for european highly renewable scenarios. In: 2017 14th International Conference on the European Energy Market (EEM), pp 1–7, DOI 10.1109/EEM.2017.7982024

Jasiński M, Najafi A, Homaee O, Kermani M, Tsaousoglou G, Leonowicz Z, Novak T (2023) Operation and planning of energy hubs under uncertainty—a review of mathematical optimization approaches. IEEE Access PP:1–1, DOI 10.1109/ACCESS.2023.3237649

Jedrzejewki P (2020) Modelling the european cross-border electricity transmission. Master's thesis, KTH School of Industrial Engineering and Management, URL `https://www.diva-portal.org/smash/get/diva2:1476768/FULLTEXT01.pdf`

Keppo I, Strubegger M (2010) Short term decisions for long term problems - the effect of foresight on model based energy systems analysis. Energy 35(5):2033–2042, DOI https://doi.org/10.1016/j.energy.2010.01.019

Khahro SF, Tabbassum K, Soomro AM, Dong L, Liao X (2014) Evaluation of wind power production prospective and weibull parameter estimation methods for Babaurband, Sindh Pakistan. Energy Conversion and Management 78:956–967, DOI https://doi.org/10.1016/j.enconman.2013.06.062

Kirschbaum S, Powilleit M, Schotte M, Özbeg F (2023) Efficient solving of time-coupled energy system milp models using a problem specific lp relaxation. pp 2774–2785, DOI 10.52202/069564-0249

Marquant JF, Mavromatidis G, Evins R, Carmeliet J (2017) Comparing different temporal dimension representations in distributed energy system design models. Energy Procedia 122:907–912, DOI https://doi.org/10.1016/j.egypro.2017.07.403

Mavrotas G, Diakoulaki D, Florios K, Georgiou P (2008) A mathematical programming framework for energy planning in services' sector buildings under uncertainty in load demand: The case of a hospital in athens. Energy Policy 36(7):2415–2429, DOI https://doi.org/10.1016/j.enpol.2008.01.011

Morais H, Kádár P, Faria P, Vale ZA, Khodr H (2010) Optimal scheduling of a renewable micro-grid in an isolated load area using mixed-integer linear programming. Renewable Energy 35(1):151–156, DOI https://doi.org/10.1016/j.renene.2009.02.031

Palma-Behnke R, Benavides C, Lanas F, Severino B, Reyes L, Llanos J, Sáez D (2013) A microgrid energy management system based on the rolling horizon strategy. IEEE Transactions on Smart Grid 4(2):996–1006, DOI 10.1109/TSG.2012.2231440

Papaefthymiou G, Kurowicka D (2009) Using copulas for modeling stochastic dependence in power system uncertainty analysis. IEEE Transactions on Power Systems 24(1):40–49, DOI 10.1109/TPWRS.2008.2004728

Parra D, Valverde L, Pino FJ, Patel MK (2019) A review on the role, cost and value of hydrogen energy systems for deep decarbonisation. Renewable and Sustainable Energy Reviews 101:279–294, DOI https://doi.org/10.1016/j.rser.2018.11.010

Pfenninger S, Staffell I (2016) Long-term patterns of european pv output using 30 years of validated hourly reanalysis and satellite data. Energy 114:1251–1265, DOI https://doi.org/10.1016/j.energy.2016.08060

Teichgraeber H, Brandt AR (2022) Time-series aggregation for the optimization of energy systems: Goals, challenges, approaches, and opportunities. Renewable and Sustainable Energy Reviews 157:111984, DOI https://doi.org/10.1016/j.rser.2021.111984

Wang C, Nehrir MH (2008) Power management of a stand-alone wind/photovoltaic/fuel cell energy system. IEEE Transactions on Energy Conversion 23(3):957–967, DOI 10.1109/TEC.2007.914200

Yilmaz HU, Mainzer K, Keles D (2020) Improving the performance of solving large scale mixed-integer energy system models by applying the fix-and-relax method. 2020 17th International Conference on the European Energy Market (EEM) pp 1–5, DOI 10.1109/EEM49802.2020.9221934

Yuan X, Chen C, Jiang M, Yuan Y (2019) Prediction interval of wind power using parameter optimized beta distribution based lstm model. Applied Soft Computing 82:105550, DOI https://doi.org/10.1016/j.asoc.2019.105550

Zipkin PH (1977) Aggregation in linear programming [electronic resource]. Ph.d. thesis, Yale University, URL `https://search.library.yale.edu/catalog/9851031`, source: Dissertation Abstracts International, Volume: 39-03, Section: B, page: 1449. Access restricted by licensing agreement.

Zipkin PH (1980) Bounds for row-aggregation in linear programming. Oper Res 28:903–916, URL `https://api.semanticscholar.org/CorpusID:207240221`

A Scenario Generation

To estimate the optimal capacities for the CEP through a stochastic approach, realistic and diverse weather scenarios are needed, so to capture the variability and uncertainty of power generation through renewable sources over extended periods. In order to generate such scenarios, samples are extracted from a joint probability density function (PDF) fit on historical data.

In our model, we used an hourly time-step ($T = 8760$) and fit the wind and solar distributions separately for each country considered.

To model the marginal probability distributions corresponding to the power output of wind turbines for each hour of the year, a Weibull distribution was used, justified by its proven effectiveness in capturing the variability and skewness of wind power distributions (Khahro et al. 2014). For solar power, Beta distributions were employed, as in Yuan et al. (2019). To fit our model, we used a dataset containing 30 years of data for various European countries, which was collected by Pfenninger and Staffell (2016). On the other hand, electricity load is taken from the ENTSO-E Statistical Reports (2024). In this simple model, while fitting on historical data we did not account for possible changes in future climate, since the focus lies mostly in the computational aspect.

To account for interdependence between temporally near time-steps, we coupled these distributions using a Gaussian Copula approach, which captures the dependencies between hourly power outputs effectively. This approach accurately represents the coupled behavior in renewable stochastic systems (Papaefthymiou and Kurowicka 2009).

A possible improvement of the generation process could be to fit wind and PV data jointly in the copula step, potentially also including load scenarios with the generation scenarios through the same approach. This would consider dependence between Energy Demand and weather conditions, but it would necessitate of the historical dataset provided for the corresponding grid, and would also further increase computational costs.

A.1 Parametric Estimation of Wind Power distribution

The parameters defining the Weibull Distribution are estimated using the Maximum Likelyhood Estimation (MLE). The Weibull density function is given by:

$$f(x;\theta,\gamma) = \left(\frac{\gamma}{\theta}\right) x^{\gamma-1} \exp\left(-\left(\frac{x}{\theta}\right)^{\gamma}\right) \tag{39}$$

where $\theta, \gamma > 0$ are the scale and shape parameters, respectively.

Given observations $X_1, \ldots, X_n$, the log-likelihood function is:

$$\log L(\theta,\gamma) = \sum_{i=1}^{n} \log f(X_i \mid \theta,\gamma) \tag{40}$$

The optimum solution is found by searching for the parameters for which the gradient is zero:

$$\frac{\partial \log L}{\partial \theta} = -\frac{n\gamma}{\theta} + \frac{\gamma}{\theta^2} \sum_{i=1}^{n} x_i^{\gamma} = 0 \tag{41}$$

Eliminating θ, we get:

$$\left[\frac{\sum_{i=1}^{n} x_i^{\gamma} \log x_i}{\sum_{i=1}^{n} x_i^{\gamma}} - \frac{1}{\gamma}\right] = \frac{1}{n} \sum_{i=1}^{n} \log x_i \tag{42}$$

This can be solved to get the MLE estimate $\hat{\gamma}$. This can be accomplished with the aid of standard iterative procedures such as the Newton-Raphson method or other numerical procedures. This is done with the aid of the package *scipy*. Once $\hat{\gamma}$ is found, $\hat{\theta}$ can be determined in terms of $\hat{\gamma}$ as:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\gamma}}\right)^{\frac{1}{\hat{\gamma}}} \tag{43}$$

A.2 Parametric Estimation of Solar Power distribution

To estimate the α and β parameters defining the Beta distribution Y, we use the Method of Moments. The mean of the random variable Y can be expressed as $\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}$ and the variance as $Var[Y] = \frac{\alpha+\beta}{(\alpha+\beta)(\alpha+\beta+1)}$. In particular by explicating β in the first equation and substituting it in the second equation we obtain that:

$$\begin{cases} \alpha = \mathbb{E}[X]\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \\ \beta = (1-\mathbb{E}[X])\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \end{cases} \tag{44}$$

By substituting the mean and the variance with their empirical approximation we obtain the Method of Moments estimator for α and β.

A.3 Parametric Copula Estimation

The cumulative density function of both the Weibull and Beta distributions are continuous and invertible. Therefore, the random variables $U_t \coloneqq F_{Y_t}(Y_t)$ have a uniform distribution over $[0,1]$. The copula of the random variables $\{Y_t\}_{t\in T}$ is defined as the function $C : [0,1]^T \to [0,1]$ such that

$$C(F_{Y_1}(y_1), \dots, F_{Y_T}(y_{|T|})) = P(Y_1 \le y_1, \dots, Y_{|T|} \le y_{|T|}). \tag{45}$$

This function always exists because of Sklar's Theorem. For a given correlation matrix Σ, the Gaussian Copula with parameter matrix Σ is defined as

$$C_\Sigma^{\mathrm{Gauss}}(u_1, \dots, u_T) \coloneqq \Phi_\Sigma(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_T)),$$

where Φ, Φ_Σ are the cumulative distribution functions of Gaussian variables having distribution $\mathcal{N}(0,1)$ and $\mathcal{N}(\mathbf{0}, \Sigma)$ respectively. In particular if $C_\Sigma^{\mathrm{Gauss}}$ is the copula associated with the random variables $\{Y_t\}_{t\in T}$ then we have that the random variables $Z_t = \Phi^{-1}(F_{Y_t}(Y_t)) = \Phi^{-1}(U_t)$ have joint distribution equal to $\mathcal{N}(0,\Sigma)$. This follows from:

$$\begin{aligned} P(Z_1 \le z_1, \dots, Z_T \le z_t) &= P(\Phi^{-1}(U_1) \le z_1, \dots, \Phi^{-1}(U_T) \le z_T) = \\ &= P(U_1 \le \Phi(z_1), \dots, U_T \le \Phi(z_T)) = \\ &= C_\Sigma^{\mathrm{Gauss}}(\Phi(z_1), \dots, \Phi(z_t)) = \\ &= \Phi_\Sigma(z_1, \dots, z_T) \end{aligned}$$

In particular, given the realization $\{y_{t,j}\}_{t\in t, s\in\mathcal{J}}$ of the variables $\{Y_t\}_{t\in T}$, an unbiased estimation of the parameter matrix Σ is the empirical covariance matrix $\hat{\Sigma}$ of the samples $\{\Phi^{-1}(\hat{F}_{Y_t}(y_{t,j}))\}_{t\in T, s\in\mathcal{J}}$, where $\hat{F}_{Y_t}$ is the estimated marginal distribution of the variable Y_t obtained as seen in Sections A.1 and A.2.

Finally, we can generate samples from a Multivariate Gaussian random variable $(Z_t, t \in T)$ having distribution $\mathcal{N}(0,\hat{\Sigma})$. Then the power output scenarios are obtained from these samples by following the previous steps backwards, that is, for each sample, computing $\hat{F}_t^{-1}(\Phi(Z_t))$ for all $t \in T$.

B Parameters

Numerous parameters characterize the grid and are incorporated into the model. The primary parameters relate to the capital costs associated with the infrastructure to be constructed. The assumed costs for photovoltaic panels and wind turbines are $cs = 400$€ and $cw = 3,000,000$€, respectively. Additionally, capital costs for fuel cells and electrolyzers are

considered. Given that hydrogen infrastructure is often developed by repurposing existing facilities, the estimation of investment costs is highly location-dependent and lies beyond the scope of this study. Therefore, for the purposes of our analysis, instead of representing the actual investment for the facilities, a minimal "symbolic" cost is assigned per unit of capacity, so that, by minimizing, the model estimates needed conversion capacities $mhte_n$ and $meth_n$.

The storage of hydrogen incurs costs that are influenced by various factors, including the capital costs of the storage technology, operating costs, and the duration for which hydrogen is held in storage. In our model, we focus solely on the parameter ch, which represents the capital cost of the storage infrastructure, to be multiplied by the maximum storage requirement nh. We do not account for marginal costs associated with maintaining hydrogen in storage.

In this model we assume there are no marginal costs associated to PV and wind power production: the operating costs of the farms throughout their life-cycle can be factored into the capital costs, and there are no additional costs linked to the production itself.

Conversely, the marginal costs of conversion within electrolyzers and power cells are relevant.

For electrolyzers, we consider the Levelised cost of hydrogen (LCOH) to account for both marginal costs and capital costs. Such cost is dependent on the country's specific market condition and can be calculated through the European Hydrogen Observatory (2023) calculator.

The parameters $fhte_n$ and $feth_n$ are defined as scalar values ranging from 0 to 1 to represent the efficiency of the conversion processes between hydrogen and electricity. It is assumed that 1 kg of hydrogen possesses an energy value of 33 kWh. Therefore, when considering an electrolyzer operating at maximum efficiency ($feth = 1$), one MWh of electricity can produce approximately $\frac{1000}{33} \simeq 30$ kg of hydrogen. For our analysis, we adopt a standard efficiency value of $feth = 0.66$, resulting in the production of 20 kg of hydrogen per MWh.

Conversely, in a fuel cell functioning at maximum efficiency ($fhte = 1$), 1 kg of hydrogen can yield 33 kWh. For our model, we utilize a value of $fhte = 0.75$, which corresponds to an output of 24.75 kWh per kg of hydrogen. It is important to note that actual efficiencies can vary significantly based on the specific technology employed. Additionally, both chemical and physical constraints currently limit achievable efficiencies to levels no higher than approximately 0.80 to 0.85 (Dawood et al. 2020).

In this model, we assume that the flow of electricity incurs no marginal cost and experiences no power loss (a more detailed modeling of these factors is beyond the scope of this project). However, we assign a cost to the use of hydrogen pipelines (or other means of transfer). The existing capacity of transmission lines and hydrogen pipelines is also set through respective parameters. Existing NTC can be estimated using the methodology outlined by Jedrzejewki (2020, chapter 5), by collecting data from Entso-e Power Statistics and Transparency Platform (2024).

Finally, the model allows for upper bounds to be placed on the variables, based on either technological and physical constraints — such as facility dimensions — or political considerations, for instance, local opposition to wind turbines.

The parameters ES, EW, EL, and HL, indexed by scenario, time-step, and node, represent the time series of power generation and load values for various scenarios at each grid node. The method used for generating these parameters is detailed in Appendix A.

A summary of all model parameters can be found in Table2.