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

Modelling of a fully renewable energy grid with hydrogen storage: time aggregation for a scalable Capacity Expansion Problem

Bianca Urso · Gabor Riccardi

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 employs a validation function, which evaluates with a Rolling Horizon method the feasibility of the aggregated solutions and selects the time interval on which the validation fails. The second method uses the proportion of net power production in each time step relative to the aggregated time interval. These selection methods are then compared against a random interval selection approach.

Keywords Electric power system · Stochastic Programming · Rolling Horizon · Time aggregation · Renewable Energy

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

B. Urso
IUSS School of Advances Studies, Palazzo del Broletto, Piazza della Vittoria, 15 – 27100 Pavia PV, Italy
Tel: +39 0382 375811
Fax: +39 0382 375899
E-mail: bianca.urso@iusspavia.it

G. Riccardi
Dipartimento di Matematica "F. Casorati" Via Adolfo Ferrata, 5 – 27100 Pavia

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 power distribution and transmission in smart grids (European Commission 2024). One of the main difficulties arising when designing an electric power system relying on renewables is the great variability of the generation of electricity through wind and solar, 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 elecrolyzers (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 though makes it impossible to plan long-term by optimizing on forecasts, and requires a statistical approach to ensure a robust model.

Up to now, common approaches have adopted Stochastic Programming (SP) or Robust Optimization (RO) models, along with hybrid models involving Information Gap Decision Theory or Chance Constraint (Jasiński et al. 2023). While initially favored, the SP approach comes with high computational burden, so RO models have seen more popularity in recent years, despite the drawback of being conservative methods with 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, grid upgrades to augment Net Transfer Capacity (NTC) and so on. Nested within the CEP is an Economic Dispatch (ED) problem concerning the operational costs of said infrastructure. The problem is well suited to be modelled through mixed integer linear programming (MILP), as is explained in detail for example in Morais et al. (2010).

The CEP for investment planning requires to look at long time horizons, and on the other hand intra-day variability in generation is the main complexity driver for the ED problem, so the time horizon must be modelled by a large quantity of fairly tight time steps. Furthermore, Large scale grids can be modelled with various degrees of spacial aggregation, as is explored by Hörsch and Brown (2017) and by Biener and Garcia Rosas (2020), and problem size increases more than linearly with the number of nodes. Thus the temporal and spacial characteristics of the model bring the MILP size to increase rapidly. This is especially demanding in the case of SP, 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 RO model to optimize the operation of a micro-grid composed of two 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 "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 to the search for integer solutions. The same method is applied by Kirschbaum et al. (2023) for the optimization of medium-scale industrial energy systems.

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

Another possible method to represent the temporal dimension in order to reduce computational costs is the "Typical Days" approach: that is, selecting a reduced number of days to act as representatives for the season, and optimizing only on them. In Domínguez-Muñoz et al. (2011) the selection of the days is carried out before optimization through a clustering technique. Marquant et al. (2017) compare the Typical Days and the RH methods in terms of solving time and accuracy on a selected distributed energy system model.

In our work, we build a 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 design of the grid, stochastically on the scenarios for the inner ED problem. A RH method is then proposed for the validation of the results for the CEP obtained in the perfect foresight optimization, rather than as a stand-alone technique, to ensure that given a solution to the CEP, the ED problem admits feasible solutions even in a limited-foresight environment. The idea is to solve the CEP as to build a grid that can be operated as a "smart grid", through a control system that optimizes on day ahead forecasts.

Further, we present our attempt at dealing with the great computational requirement by aggregating on time steps to reduce the model size. The optimization is carried out through an iterative procedure gradually refining the partition of the time horizon, and converging to the optimal solution. The end result is an optimization on Typical Days, but the selection of the days happens within an iterative process rather than before the start of optimization. To guide the selection of these progressively tighter relaxations of the perfect foresight model, two methods are discussed and compared: a first one making use of the aforementioned RH validation function, and a second one devised ad hoc based on the problem structure.

This paper is organized as follows.

Section 2 descibes the formulation of the LP problem: the first subsection details the perfect foresight model, and the second adapts the structure from the first to the RH approach.

Section 3 deals with time aggregation. In subsection 3.1 the LP problem associate to the variables aggregated over time is defined, and is shown to be a relaxation of the disaggregated one; In subsection 3.2 a method is proposed to iterate through progressively finer time aggregations, in order toconverge to the optimal perfect foresight solution. The use of the RH method defined in the previous section is proposed to select the interval to disaggregate. In subsection 3.3, the structure of the constraint matrix is exploited to derive conditions under which the aggregated solution is feasible for the disaggregated problem, and the result is used to devise a heuristic for the disaggregation selection.

Finally, section 4 contains the computational results: the three aggregation-iteration methods are compared over time and number of iterations for convergence.

2 MODEL

2.1 LP Formulation

Our model describes a network in Europe that is to be powered and supplied of hydrogen trough power generated by photovoltaic panels and wind turbines, converted to hydrogen through electrolysis and potentially reconverted in fuel cells.

The network is represented by an undirected multigraph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ corresponds to the nodes in the network, and $\mathcal{E} = \mathcal{E}_H \cup \mathcal{E}_P$ represents transmission lines ($e \in \mathcal{E}_P$) and hydrogen lines ($e \in \mathcal{E}_H$). Each of these nodes can be in different countries in Europe, and the power generated by wind and solar power depends on the node location. In particular, if node $n \in \mathcal{N}$ is in France, the scenarios for power generation at n will be generated using parameters fitted to France's data.

Each node has its generators, hydrogen storage, fuel cells and hydrolyzers, for which the capacity is to be decided. Likewise, the CEP aims to solve for transmission line NTC and hydrogen pipe transmission capacity for each edge of the network. The basic formulation of the LP described in this subsection allows us to solve the CEP with perfect foresight.

We decided to model our problem as a LP problem instead of MILP, as would be standard in the literature. This is because when modelling a large grid with high demand, optimizing on continuous variables and then rounding up to the closest integer for variables regarding the number of PV panels and turbines notably decreases computational costs, without relevant difference in the final cost. The same wouldn't be possible when optimizing over micro-grids.

The model takes in input the generation and load scenarios of the given area along with various parameters indicating costs and efficiency of the current state of technology and possible upper bounds for the decision variables. The CEP and the ED problems for all scenarios are solved concurrently. When optimizing over multiple scenarios jointly, the solver returns the minimal amount of

infrastructure and capacities that is needed to have feasibility (that is, demand met at all time and no blackouts) over all scenarios in input, with minimal cost. Cost is considered to be the sum of capital costs and average operational costs of the infrastructure over the scenarios.

For the rest of this article, when talking about a solution to the LP problem, we refer to a realization of variables solving the CEP and ED problem jointly. When mentioning a solution to the CEP problem, we refer to those components of the corresponding LP problem solution that are not time or scenario dependent. When talking about a solution to the ED problem only, we refer to a realization of the time and scenario dependent variables that solve the problem for a given *fixed* CEP solution.

The optimization problem is solved using the Gurobi solver (Gurobi Optimization, LLC 2024).

2.1.1 Decision Variables

The main variables that are of interest to the policy maker are, for each location n, the number of wind turbines and PV panels to install, as well as hydrogen storage capacity. Stored hydrogen is considered to be the total of liquid and gas hydrogen to be stored. Our model does not assume a distinction between the two forms, and considers hydrogen to be immediately ready for long-term storage as soon as it is converted from electricity, as well as instantaneously convertible to electricity in fuel cells at need.

Important values to consider when planning the grid are the power capacity and conversion speed of fuel cells and electrolyzers: variables $mhte_n$ and $meth_n$, indexed by node n, indicate for each location the maximum amount of energy that can be converted at a single time step respectively from hydrogen to electricity and vice versa. These values are essential to estimate in order to design a grid that can effectively accommodate peak production and supply during low production periods.

Transmission on the grid lines is considered to be bound by Net Transfer Capacity (NTC). Existing NTC can be estimated through the procedure followed by Jedrzejewki (2020, chapter 5), by collecting data from Entso-e Power Statistics and Transparency Platform (2024). Improvements on the existing capacity for power lines or hydrogen transport infrastructure are considered through variables $addNTC_l$ and $addMH_l$.

Variables linked to the inner ED problem are indexed by scenario $j \in J$, time step $t \in \{1...T\}$, and either node $n \in \mathcal{N}$ or edge $l \in \mathcal{E}_H$ or $l \in \mathcal{E}_P$. For the variables pertaining to transmission on an edge, two distinct variables are considered, one for each direction. This way, all variables are set to be non-negative. This will be relevant for the formulation of the time-aggregated relaxation of the LP problem.

See Table 1 for the summary of all decision variables.

Table 1 Decision variables

Name	Unit	Description
ns_n	-	Number of solar units at node $n \in \mathcal{N}$
nw_n	-	Number of wind units at node $n \in \mathcal{N}$
nh_n	kg	Storage capacity at node $n \in \mathcal{N}$
mhte_n	kg	Maximum hydrogen to electricity conversion capacity (per time step) at node $n \in \mathcal{N}$
meth_n	MWh	Maximum electricity to hydrogen conversion capacity (per time step) at node $n \in \mathcal{N}$
addNTC_l	MWh	Additional net transfer capacity on line $l \in \mathcal{E}_P$
addMH_l	kg	Additional hydrogen transfer capacity on pipe $l \in \mathcal{E}_H$
$\text{H}_{j,t,n}$	kg	Stored hydrogen at node n, time t, scenario j
$\text{HtE}_{j,t,n}$	kg	Hydrogen converted to electricity at time t, scenario j
$\text{EtH}_{j,t,n}$	MWh	Electricity converted to hydrogen at time t, scenario j
$\text{P\_edge}^{+}_{j,t,l}$	MWh	Power passing through line l at time t, scenario j
$\text{P\_edge}^{-}_{j,t,l}$	MWh	Power passing through line l at time t, scenario j
$\text{H\_edge}^{+}_{j,t,l}$	kg	Hydrogen transported on line l at time t, scenario j
$\text{H\_edge}^{-}_{j,t,l}$	kg	Hydrogen transported on line l at time t, scenario j

2.1.2 Parameters

There are numerous parameters that describe the grid and are passed to the model. The main ones are related to capital costs of the infrastructure to be built and the following values are assumed for panels and turbines: $cs = 400$€, $cw = 3000000$€. We also set capital costs for fuel cells and electrolyzers: since hydrogen infrastructure is usually obtained by reconverting existing infrastructure from other purposes, the estimation of investment costs is very location dependent and beyond the scope of this work. Thus for our purposes, instead of representing the actual investment for the facilities, a minimal "symbolic" cost is assigned per unit of capacity, so that in minimizing the model estimates needed conversion capacities $mhte_n$ and $meth_n$.

The storage of the hydrogen has a cost that depends on various factors: capital cost of the technology used for storage, operating costs, length of time that the hydrogen is kept in storage. For our model we only set the parameters ch, to be multiplied by the maximum storage needed (nh), representing capital cost of storage infrastructure. We ignore marginal costs of keeping the hydrogen stored.

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

Parameters $fhte_n$ and $feth_n$ are set as scalars between 0 and 1 to indicate efficiency of the conversion from hydrogen to electricity and vice versa. It is assumed that 1kg of hydrogen has an energy value of 33kWh. Thus if we consider an electrolyzer operating at maximum efficiency ($feth = 1$), one MWh of electricity yields $1000/33 \simeq 30$kg of hydrogen. For our purposes, a standard value of $feth = 0.66$ is considered, thus 1MWh yields 20kg of hydrogen. Conversely, in a fuel cell operating at maximum efficiency ($fhte = 1$) 1kg of hydrogen yields 33kWh. We consider a value of $fhte = 0.75$, yielding 24.75kWh per kg of hydrogen. Actual efficiencies vary a lot depending on the technology used. Furthermore, chemical and physical constraints make it so that efficiencies higher than 0.80-0.85 are currently considered unachievable (Dawood et al. 2020).

Additionally, we assume the flow of electricity has no marginal cost nor power loss (the modelling of that problem is beyond the scope of this project), whereas we do set a cost for the use of hydrogen pipes (or other means of transfer). The existing capacity of transmission lines and hydrogen pipes is also set.

Finally, the model allows for upper bounds to be placed on the variables, based on either technological and physical constraints (dimension of the facilities) or because of political choices (e.g. local population unfavourable to wind turbines).

The parameters ES, EW, EL, HL, indexed by scenario, time step and node represent the time series of power generation and load values for different scenarios in every node of the grid. The method used for generation is described in Appendix A.

See Table 2 for the summary of all parameters.

2.1.3 Objective Function

The cost function is given by the sum of all capital costs of installing infrastructure, plus all marginal costs of the hydrogen to electricity and electricity to hydrogen conversions and hydrogen transfer on the edges at each time step.

Table 2 Model parameters

Name	Unit	Description
cs	€	Cost of one Solar Panel at node n
cw	€	Cost of one Wind Turbine at node n
chte_n	€/kg	Conversion cost of hydrogen to electricity
ceth_n	€/MWh	Conversion cost of electricity to hydrogen
ch_n	€/kg	Cost of hydrogen storage capacity
cH_edge_l	€	Cost of transferring 1kg of hydrogen on edge $l \in \mathcal{E}_H$
cNTC_l	€/MWh	Cost of adding NTC to line $l \in \mathcal{E}_P$
cMH_l	€/kg	Cost of adding H_2 transfer capacity to line $l \in \mathcal{E}_H$
cmhte	€/kg	Cost of needed HtE capacity per unit
cmeth	€/MWh	Cost of needed EtH capacity per unit
fhte_n	-	Efficiency of hydrogen to electricity conversion
feth_n	-	Efficiency of electricity to hydrogen conversion
NTC_l	MWh	Net Transfer Capacity on line $l \in \mathcal{E}_P$
MH_l	kg	Hydrogen transfer capacity on edge $l \in \mathcal{E}_H$
Mns_n	-	Maximum number of solar panels at node n
Mnw_n	-	Maximum number of wind turbines at node n
Mnh_n	kg	Maximum hydrogen storage capacity
Mhte_n	kg	Upper bound for $mhte$
Meth_n	MWh	Upper bound for $meth$
$\text{ES}_{j,t,n}$	MWh	Power output of a single solar panel
$\text{EW}_{j,t,n}$	MWh	Power output of a single wind turbine
$\text{EL}_{j,t,n}$	MWh	Electricity load
$\text{HL}_{j,t,n}$	kg	Hydrogen load

Let d be the number of scenarios, and T the number of time steps. The objective function is as follows:

$$
\begin{aligned}
\min \quad & \sum_{k\in\mathcal{N}} (\text{cs}_k \cdot \text{ns}_k + \ \text{cw}_k \cdot \text{nw}_k + \text{ch}_k \cdot \text{nh}_k) \ + \\
+ \quad & \sum_{k\in\mathcal{N}} (\text{cmhte}_k \cdot \text{mhte}_k + \text{cmeth}_k \cdot \text{meth}_k) \ + \\
+ \quad & \sum_{l\in\mathcal{E}_P} (\text{cNTC}_l \cdot \text{addNTC}_l) + \sum_{l\in\mathcal{E}_H} (\text{cMH}_l \cdot \text{addMH}_l) \ + \\
+ \quad & \frac{1}{d} \sum_{j=1}^{d} \sum_{t=1}^{T} \Bigg(\sum_{k\in\mathcal{N}} (\ \text{ch\_t}_k \cdot \text{H}_{j,t,k} + \text{chte}_k \cdot \text{HtE}_{j,t,k} + \text{ceth}_k \cdot \text{EtH}_{j,t,k}) \ + \\
& \qquad + \sum_{l\in\mathcal{E}_H} (\text{cH\_edge}_l \cdot \text{H\_edge}_{j,t,l}) \Bigg)
\end{aligned}
\tag{1}
$$

The $1/d$ factor in front of the marginal costs allows to average over the scenarios, whereas the capital costs are the same for all scenarios. Thus, ignoring the costs of $mhte_k$ and $meth_k$, the objective function value gives an estimate of the actual costs (in €) for the set up and maintenance of the system throughout the length of the time horizon.

2.1.4 Constraints

The following constraints are to ensure that for all time steps t and all scenarios j, the electricity load and the hydrogen load are met. The measure units are MWh and kg respectively, and conversion factors are considered for HtE and EtH respecively. Let $Out(n)$ and $In(n)$ indicate the outgoing and incoming edges from node n on the respective graph. For simplicity, we indicate with $P\_edge_{j,t,l}$ the difference of the corresponding variables $P\_edge^{+}_{j,t,l} - P\_edge^{-}_{j,t,l}$, and analogously for H_edge. When solving, it is sufficient to assign a symbolic cost to them to ensure that only one of them is non zero at each time step. Then for each node n, scenario j and time step t, the following flow balance constraints are imposed:

$$\begin{aligned} \text{Electricity Balance:} \quad & \mathrm{ns}_n \cdot \mathrm{ES}_{j,t,n} + \mathrm{nw}_k \cdot \mathrm{EW}_{j,t,n} - \mathrm{EL}_{j,t,n} + \\ & + 0.033 \quad \cdot fhte_k \cdot \mathrm{HtE}_{j,t,n} - \mathrm{EtH}_{j,t,n} + \\ & + \sum_{l \in Out(n)} \mathrm{P\_edge}_{j,t,l} + \sum_{l \in In(n)} \mathrm{P\_edge}_{j,t,l} \geq 0; \end{aligned} \tag{2}$$

$$\begin{aligned} \text{Hydrogen Storage:} \quad \mathrm{H}_{j,t+1,n} = \; & \mathrm{H}_{j,t,n} - \mathrm{HL}_{j,t,n} + \\ & + 30 \cdot feth_k \cdot \mathrm{EtH}_{j,t,n} - \mathrm{HtE}_{j,t,n} + \\ & - \sum_{l \in Out(n)} \mathrm{H\_edge}_{j,t,l} + \sum_{l \in In(n)} \mathrm{H\_edge}_{j,t,l} \end{aligned} \tag{3}$$

We ask that the consumed electricity be less or equal than the produced or received electricity at all times. On the grid itself, the two sides should be equal, but we observe that $ns \cdot ES_{j,t} + nw \cdot EW_{j,t}$ indicate the maximum power that can be generated with set weather conditions, whereas actual production will be regulated to meet demand through curtailment.

The stored hydrogen at time $t+1$ is the result of what was stored at time t adjusted by what was converted and what was sent to the industrial load. For $t = T$ (the last time step) we set the same constraint on hydrogen by considering $t+1$ to be index 1: this way we avoid placing a "start time" at an arbitrary place within the year (time is rendered modulo the year) and we avoid the model asking for conveniently high initial storage values of hydrogen appearing out of thin air.

The total storage and conversion capacities are calculated by minimizing the maximum over time and scenarios of the variables $H_{j,t}$, $EtH_{j,t}$ and $HtE_{j,t}$,

for all scenarios j, time steps t and nodes n:

$$\text{Storage Capacity Limit:} \quad \mathrm{H}_{j,t,n} \leq \mathrm{nh}_n; \tag{4}$$

$$\text{EtH Conversion Limit:} \quad \mathrm{EtH}_{j,t,n} \leq \mathrm{meth}_n; \tag{5}$$

$$\text{HtE Conversion Limit:} \quad \mathrm{HtE}_{j,t,n} \leq \mathrm{mhte}_n. \tag{6}$$

Finally, edge capacities on the respective graphs are considered for all scenarios j, time steps t and nodes n:

$$\text{Net Transfer Capacity:} \quad \mathrm{P\_edge}^{\pm}_{j,t,l} \leq \mathrm{NTC}_l + \mathrm{addNTC}_l; \tag{7}$$

$$\text{Hydrogen Transfer Capacity:} \quad \mathrm{H\_edge}^{\pm}_{j,t,l} \leq \mathrm{MH}_l + \mathrm{addMH}_l. \tag{8}$$

2.2 Validation: Rolling Horizon

While computing the optimal solution on a batch of scenarios by solving the LP model described in section 2.1, the solver "knows the future" for those scenarios. That is, the criteria it uses to determine, at each time step, how much electricity or hydrogen should be converted or transmitted, and where, is a mathematical minimum that is informed by the knowledge of what is needed at any time during the one year scenario. In real life, accurate forecasts for weather, and consequently for power generation, are known on a day ahead basis, at most two days ahead. We are thus interested in evaluating whether an optimal grid as given by the solution of the CEP problem on a batch of train scenarios can (1) be operated without knowledge of the future in order to satisfy demand on the same scenarios it was trained on, and (2) generalize to new test scenarios.

Let's first consider the case where the grid is composed of a single node. The actions and choices of a power grid administrator of such an isolated micro-grid are very much limited to "given extra energy, store it, up to storage limit". Such a strategy is for example discussed in more detail in Wang and Nehrir (2008). A deterministic system control can be easily designed to check whether a single node CEP solution is sufficient for feasibility over a certain scenario, even without day ahead forecasts. This is not true anymore once the node is taken out of isolation: at each moment, without knowledge of the full future, each fictional administrator at each node must choose whether to store the energy generated at their node, whether to send it to a neighbor (and how much to which neighbor) or how to collect missing energy to match their node's demand.

In order to operate a multi-node grid, we propose to use the Rolling Horizon optimization technique. The basic idea is to divide the time horizon into smaller periods and to progressively optimize on each period, passing the variables of the solved periods as fixed to the next. In our case, we are interested in periods of length one day each, representing the knowledge of future that is given to the grid administrator by weather forecasts.

Recall that when optimizing with the LP model described in section 2.1, the solutions to the CEP and to the ED problem for all train scenarios $j \in J_{train}$ are given concurrently. Consider now a test scenario with generation and load time series $ES^{test}_{t,n}, EW^{test}_{t,n}$, $EL^{test}_{t,n}, HL^{test}_{t,n}$ (the test scenario can be in J_{train} or not).

The RH algorithm is as follows:

- Start with $H^{test}_{0,n} = \max_{j \in J_{train}} H_{j,0,n}$.
- For each day in the time horizon:
  - Optimize the inner ED problem for that day. If the problem is infeasible, break.
  - Set the hydrogen storage levels of the last time step of the day equal to the ones for the first time step of the next day.

The daily ED problem is formulated in a similar way to the LP problem described in section 2.1. The differences are that the CEP variables (ns, nh,...) are fixed, and the hydrogen storage level doesn't loop as it did in the year, but it connects to the following day.

Definition 1 (RH-feasibility) Given a solution $\mathbf{x}_{CEP}$ to the CEP solved over train scenarios $j \in J_{train}$ and a test scenario j, we consider $\mathbf{x}_{CEP}$ to be RH-feasible over scenario j if the RH optimization algorithms terminates at the end of the year and $H_{j,T,n} \geq H_{j,0,n}$ for all nodes $n \in \mathcal{N}$.

We require the last condition to mean that the storage levels at the end of the year are at least as high as they were at the start, to flag as unfeasible solutions that manage to satisfy demand throughout the year only through a net consumption of unproduced hydrogen.

Written as such, the daily optimization would tend to avoid storing hydrogen unless needed within the same day, since operating the elecrolyzers has a cost. This easily renders infeasible scenarios that would be feasible with better storage management. To solve this problem, one can introduce a loss function in the model: for example one can assign a cost to the difference between the hydrogen storage level at time step t and the average over the corresponding variables in the optimal solutions from train scenarios. Thus one defines positive variables $loss_{t,n}$ for each time step t and node n, with posiitive cost, and adds the contraint:

$$loss_{t,n} \;\geq\; \frac{1}{d}\left(\sum_{j \in J_{train}} H_{j,t,n}\right) - H^{test}_{t,n} \tag{9}$$

Another option would be to assign a slight negative cost to $H^{test}_{t,n}$, to incentivize filling up the storage, but this can inflate the estimated cost of operating electrolyzers more than necessary.

We observe that the overall solution to the ED problem throughout the full time horizon obtained by means of the RH method is not necessarily optimal, neither with nor without the added loss function. However, some results can

be obtained regarding the distance of such solution from the optimal of the ED problem. Indeed, considering the result by 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 $mhte_n$.

For a solution to the ED problem that is closer still to the perfect foresight optimal, more refined RH tecniques can be used, such as optimizing on two-day forecasts with a daily refresh rate. However, for our purposes, being able to check for feasibility with less than optimal management is sufficient.

3 Time Resolution

The scenarios generated from our gathered data 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, making 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 will be similar to neighboring days. Yet, simply considering a sample of days for each season might undermine long-term storage capacity representation. Thus we are interested in finding more efficient ways to deal with the time dimension in our problem.

3.1 Time aggregation as model relaxation

We introduce the following concept:

Definition 2 (Time partition) Given an initial time horizon $\mathcal{T} = \{1, \dots, T\}$, a time partition $P = \{I_1, ..., I_{T'}\}$ is a partition of $\mathcal{T}$ such that all subsets are intervals. Furthermore, we say that a time partition P' is finer than P if for every $I' \in P'$, there exists some $I \in P$ such that $I' \subset I$.

Given a time partition P, we can consider the problem LP_P associated to the model obtained by considering each interval in P as a single time step. For every $I \in P$, define:

$$ES_{j,I,n} \coloneqq \sum_{i \in I} ES_{j,i,n}, \qquad EW_{j,I,n} \coloneqq \sum_{i \in I} EW_{j,i,n}, \tag{10}$$

and analogously for $EL_{j,I,n}$ and $HL_{j,i,n}$.

It is easy to show that the optimal value to the aggregated problem LP_P is a lower bound for the original problem $\mathrm{LP}_{\mathcal{T}}$. Indeed, given a feasible solution of the latter, we can obtain a solution of the former by fixing the capital infrastructure variables to be the same as in $\mathrm{LP}_{\mathcal{T}}$ and letting all time dependent variables for the inner ED problems be defined as follows:

$$EtH_{j,I,n} = \sum_{i \in I} EtH_{j,i,e}, \qquad HtE_{j,I,e} = \sum_{i \in I} HtE_{j,i,e}. \tag{11}$$

Similarly we sum over $\Delta H_{j,i,e}$ to obtain $\Delta H_{j,I,e}$, and we get $P\_edge^{\pm}_{j,I,e}$ and $H\_edge^{\pm}_{j,I,e}$ by summing over $P\_edge^{\pm}_{j,i,e}$ and $H\_edge^{\pm}_{j,i,e}$ respectively, separately on the two directions. By defining the aggregated variables this way, the aggregated constraints are not violated, thus we get a feasible solution to LP_P. Observe that all variables with non zero cost are defined as greater or equal to zero, so summing over them we define a cost-preserving linear map from the solution space of $\mathrm{LP}_{\mathcal{T}}$ to the solution space of LP_P.

The above discussion holds in general in the case of any two partitions P and P' where P' is finer than P. We can summarize the above in the following observation:

Observation 3.1 *Let P and P' be two time partitions of $\mathcal{T} = \{1...T\}$ such that P' is finer than P. Let $V_P \subset \mathbb{R}^{N_P}$ and $V_{P'} \subset \mathbb{R}^{N_{P'}}$ be the spaces of feasible solutions of LP_P and $LP_{P'}$, respectively. Then there exists a linear map $L : \mathbb{R}^{N_{P'}} \to \mathbb{R}^{N_P}$ such that $L(V_{P'}) \subset V_P$ and $c_P(L(x)) = c_{P'}(x)$, where c_P is the cost function of LP_P and $c_{P'}$ is the cost function of $LP_{P'}$.*

From this observation follows that:

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

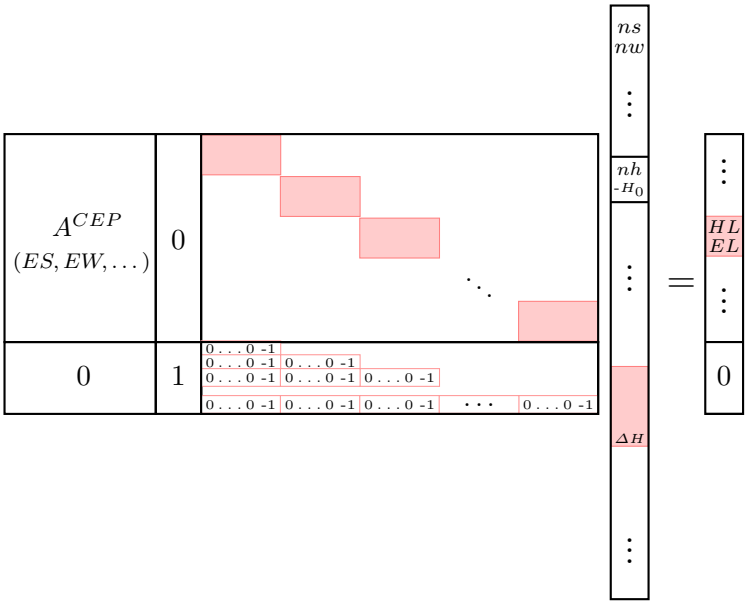
To clarify what time aggregation implies on the model constraints, we express the LP problem in standard matrix form (as expressed in formulation 12), in order to highlight its inner structure.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \mathbf{c}^\intercal \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \\ & x \geq 0 \end{aligned} \tag{12}$$

Fistly, we can reformulate the model described in subsection 2.1 by introducing a new set of variables $\Delta H_{j,t,n} \coloneqq H_{j,t+1,n} - H_{j,t,n}$, replacing the original variables $H_{j,t,n}$. We add variables $H_{j,0,n}$ to represent initial storage conditions. Constraints (3) and (4) shall be reformulated accordingly, with the latter becoming:

$$H_{j,0,n} + \sum_{i=1}^{t} \Delta H_{i,j,n} \leq n h_n, \quad \forall\, t = 1...T, \tag{4'}$$

In the problem under consideration, we have various types of constraints: Electricity Balance (2), Hydrogen Balance (3'), Hydrogen Storage (4′), maximum capacity of the time dependent variables (5),(6),(7),(8), and bounds on the CEP variables. By splitting the variable vector $\mathbf{x}$ into $\mathbf{x}^{CEP}$ (excluding nh and $H_{j,0,n}$, that are treated individually), and $\mathbf{x}^{ED}$, we can view the constraint matrix A as divided in the following sections:



(13)

The upper rows contain the balance constraints regarding electricity and hydrogen, and bound constraints for the time dependent variables (whose bounds are CEP variables). The bottom section of the matrix accounts for storage constraints (4′): summing over $\Delta H_{j,i,n}$ returns the original variable $H_{j,i,n}$, bound by nh_n. The handling of the storage is here represented for a single node and scenario. All other variable bounds are not represented, since they pertain to one variable at the time and don't really influence the overall reasoning. Slack variables to transform inequality constraints to equality are omitted.

For clarity, we represent now a simplified version of the rows corresponding to a single time step, ignoring edge variables and including only balance constraints and the upper bound for HtE as an example.

$$\left[\begin{array}{ccc|ccc} ES_{j,t,n} & EW_{j,t,n} & 0 & 0.033 \cdot fhte_n & -1 & 0 \\ 0 & 0 & 0 & -1 & 30 \cdot feth_n & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \text{ns} \\ \text{nw} \\ \text{mhte}_n \\ \hline \vdots \\ HtE_{j,t,n} \\ EtH_{j,t,n} \\ \Delta H_{j,t,n} \\ \vdots \end{array}\right] \begin{array}{c} \geq \\ = \\ \geq \end{array} \left[\begin{array}{c} EL_{j,t,n} \\ HL_{j,t,n} \\ 0 \end{array}\right] \tag{14}$$

We observe that the coefficients within the blocks are the same for all blocks, and don't depend on time or scenario.

Consider now the upper part of matrix 13. By summing over the respective rows corresponding to the time steps within the same interval I in a time partition P, we obtain the summations in 10. Row operations such as this always yield LP relaxations. In order to trace back to the model with aggregated variables as well, as defined in 11, a column operation must be carried on. In general, column operations (especially when reducting the dimension of the variable space) don't yield problem relaxations. However, thanks to the block structure with costant coefficients in all the blocks, the upper part of the contraint matrix of the aggregated LP formulation is indeed a relaxation of the disaggregated one. More specifically, observe that the single block within the constraint matrix of the aggregated LP_P is exactly the same as the blocks in the disaggregated matrix. A formal proof of this fact, written for more general row and column operations, is given in Appendix B.

The bottom part of the matrix is dealt with more easily: aggregating simply corresponds to dropping all rows constraining the storage for instants inside the aggregated time steps, while keeping the ones corresponding to the start/end of the aggregated time steps unchanged. Thus obtaining a relaxation.

3.2 Iteration on time partitions

Given that any time-aggregated LP is a relaxation for the original LP, but with many fewer variables, and thus lower optimization time, we wish to utilize purposely chosen aggregations to warm start the solver iteratively, converging to the optimal solution through progressively finer aggregations.

Power generation and electricity load data tend to follow very strong seasonal and daily patterns, and on a minor extent, a weekly trend. Generally, long term patterns are accounted for even with more loose aggregations. On the other hand, a grid described by the solution to the CEP_P for a loose partition will likely be unable to deal with daily variability. Indeed, it is quite common to have days with overall greater power production than load, but with peak production at noon and most of the power load during the late evening: such a discrepancy is overlooked by the aggregated model. Choosing a finer partition with well distributed medium-sized time steps (for example weekly or daily) would not solve this blind spot. For this reason, in choosing the refinement strategy for the time partitions we opted to disaggregate one single whole day per iteration into hourly time steps.

The method devised is the following:

1. Set up the model environment with enough variables for the iterations to come. Impose the constraints relative to an initial time partition, and solve.
2. Select a day with some given selection method.
3. Add the constraints relative to each hour of that day. Solve with warm start.
4. Repeat step 2 and 3 until some halting condition.

The base implementation we test considers a random selection method, and a very arbitrary halting condition on the number of iterations. This will be the baseline to compare the other methods we propose.

The first method we propose makes use of the RH validation function we discussed in subsection 2.2. The RH will optimize day-by-day with hourly time steps, that is, limited foresight on the finest time resolution. To choose the day to disaggregate, let the RH start on one of the training scenarios with the fixed $\mathbf{x}_{CEP}$ values given by the solution to the previous iteration. If the RH iteration breaks before reaching the end of the year, choose the day it fails on as the day to disaggregate. Proceed to solve the full LP problem on the new aggregation, and start validating again with RH beginning on the day it had previously failed on. If the RH makes it through the year, then the considered $\mathbf{x}_{CEP}$ values are part of a feasible solution for the LP problem on the whole horizon, on the finest time resolution, thus we have solved the original problem we were aiming for.

One advantage of using the RH validation function within the iterative aggregation method is that it automatically provides an effective halting method that guarantees feasibility for the original problem.

Note that even though the final $\mathbf{x}_{CEP}$ comes from the optimal solution of a relaxation, this is not enough to guarantee optimality for the finer resolution CEP, since marginal costs of conversion to and back from hydrogen, and of its transportation, aren't accounted for fully.

Unfortunately, a great disadvantage of using the RH validation within the iteration loop is that the computational costs it brings along is not insignificant, needing careful consideration on whether it is in effect preferable to directly optimizing on the full problem.

3.3 Rho

We now discuss a second method we propose for the selection of the day to disaggregate in step 2 of algorithm described in subsection 3.2. The goal is to design a heuristic to measure "how distant the solution to the aggregated LP is from being extended to a solution that is feasible to the disaggregated LP".

Consider a time partition P of the time horizon $\mathcal{T}$ and let $I \in P$ and $t \in I$. Let the un-aggregated $\text{LP}_{\mathcal{T}}$ problem be represented by constraint matrix A and vector $\mathbf{b}$, and analogously let $\tilde{A}$ and $\tilde{b}$ represent the aggregated LP_P. Let be a solution to LP_P.

Obviously extending $\tilde{\mathbf{x}}$ in order to immerse it in the original $\text{LP}_{\mathcal{T}}$ solution space requires to set the value to all variables relative to time steps in the interior of the aggregated steps, in a way that satisfies all respective constraints, and this is not always possible. But we can attempt to do so under very specific assumptions.

Definition 3 (Rho) Let r be a single row in matrix A, associated to a constraint regarding time step t within the upper part of matrix 13. Let R be the row corresponding to the respective aggregated constraint in matrix $\tilde{A}$.

Assume $\tilde{b}_R - \tilde{A}_R^{CEP} \cdot \tilde{\mathbf{x}}^{CEP} \neq 0$, define:

$$\rho_r \coloneqq \frac{b_r - A_r^{CEP} \cdot \tilde{\mathbf{x}}^{CEP}}{\tilde{b}_R - \tilde{A}_R^{CEP} \cdot \tilde{\mathbf{x}}^{CEP}} \ .$$

In essence, ρ_r represents, for each time step $t \in I \in \mathcal{T}$, the ratio between the net energy production at time t and the net energy production over the interval T.

Observe that summing ρ_r over all indexes r corrisponding to the same aggregated time step constraint R is equal to 1, since aggregated constraints are defined through summations (10).

Proposition 2 *With the notation introduced above, assume the following:*

- $\tilde{b}_R - \tilde{A}_R^{CEP} \cdot \tilde{\mathbf{x}}^{CEP} \neq 0$ *for all rows* R *of the upper part of* $\tilde{A}$*;*
- ρ_r *is constant over all rows* r *representing constraints for the same time step* t *(within the upper part of* A*). Denote this as* ρ_t*.*
- *for every aggregated step* $I \in P$*,* $\rho_t \geq 0$ *for all* $t \in I$*.*

Define $\mathbf{x}$ *in the solution space of* $LP_{\mathcal{T}}$ *by setting*

$$\mathbf{x}^{CEP} = \tilde{\mathbf{x}}^{CEP}, \quad \textit{and} \quad x_t^{ED} = \rho_t \cdot \tilde{x}_I^{ED} \quad \forall t \in I, I \in P.$$

Then $\mathbf{x}$ *is a feasible solution for* $LP_{\mathcal{T}}$*.*

Proof The third condition ensures all variables $\mathbf{x}^{ED}$ are positive at all times.

Let Q indicate the "block" submatrix from matrix 13. For all rows r in the upper part of matrix 13, let Q_r denote the row of Q involved in row A_r; recall that $Q_r = Q_R$. Then the following holds:

$$\begin{aligned} A_r \cdot \mathbf{x} &= A_r^{CEP} \cdot \mathbf{x}^{CEP} + Q_r \cdot \mathbf{x}_t^{ED} = \\ &= A_r^{CEP} \cdot \mathbf{x}^{CEP} + Q_R \cdot \rho_t \cdot \tilde{\mathbf{x}}_I^{ED} = \\ &= A_r^{CEP} \cdot \mathbf{x}^{CEP} + \frac{b_r - A_r^{CEP} \cdot \tilde{\mathbf{x}}^{CEP}}{\tilde{b}_R - \tilde{A}_R^{CEP} \cdot \tilde{\mathbf{x}}^{CEP}} \cdot Q_R \cdot \tilde{\mathbf{x}}_I^{ED} = \\ &= A_r^{CEP} \cdot \mathbf{x}^{CEP} + b_r - A_r^{CEP} \cdot \tilde{\mathbf{x}}^{CEP} = b_r \end{aligned}$$

For the lower part of matrix 13, the constraint reduces to:

$$\begin{aligned} \sum_{t \in \mathcal{T}} \Delta H_{j,t,n} &= \sum_{I \in P} \left(\sum_{t \in I} \Delta H_{j,t,n} \right) = \sum_{I \in P} \left(\sum_{t \in I} \rho_t \cdot \Delta H_{j,I,n} \right) = \\ &= \sum_{I \in P} \left(\sum_{t \in I} \rho_t \right) \cdot \Delta H_{j,I,n} = \sum_{I \in P} \Delta H_{j,I,n} \leq \ nh_n - H_{j,0,n} \end{aligned}$$

□

Thus, under these assumptions, we are able to construct a feasible solution for $\text{LP}_{\mathcal{T}}$ starting from $\tilde{\mathbf{x}}$, building $\mathbf{x}$ by appropriately scaling the variables $\tilde{\mathbf{x}}$ within each aggregated time step, by a factor ρ_t.

The condition on ρ_r is obviously rarely satisfied, since it would require for the power generation and load time series to be perfectly aligned (or perfectly opposed) with hydrogen load series, and across all nodes in the grid. However, we are interested in exactly those cases where generation and load are most misalligned, with the expectation that those are the cases that cause CEP_P solutions to be infeasible for $\text{LP}_{\mathcal{T}}$.

In conclusion, within the iterative procedure for refining the solution of the aggregated problem 3.2, we can select the time interval to be refined based on the extent to which it violates the conditions of proposition 2. Such cases correspond to those intervals $I \in P$ where ρ_r has greatest variance over the rows constraining the same time moment $t \in I$.

This same approach, generalized to a wider family of aggregations LP problems, is detailed in appendix B.

4 COMPUTATIONAL RESULTS

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

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 ρ-variance as defined in subsection 3.3, and (3) selecting the interval where the RH validation function discussed in 2.2 fails.

Plot 1 illustrates the cost variation at each iteration using the ρ selection method compared to random interval selection. The ρ selection method demonstrates a faster increase in cost than the random selection method, with comparable optimization times: 174 seconds for the ρ selection method and 155 seconds for the random selection method over 10 iterations.

In plot 2, we compare the cost variation at each iteration using the validation function for interval refinement against the random interval selection method. The results indicate that the former method yields a significantly faster increase, implying quicker convergence to the optimal solution. However, this approach incurs greater computational time: many seconds s.

It is worth noting that while both the ρ and random iteration methods continue up until reaching the maximum iteration limit, the validation function 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. However such adjustments

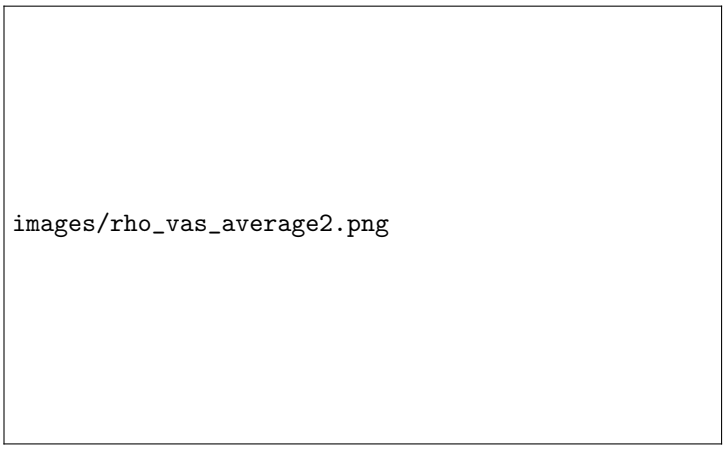


Fig. 1 Cost variation at each iteration using the ρ selection method compared to random interval selection.

Fig. 2 Cost variation at each iteration using the validation function for interval refinement against the random interval selection method.

can be made without further increasing the number of disaggregated days, keeping the problem size limited.

Figure 3 shows the optimization time across iterations. We observe a generally constant, yet slightly decreasing trend in optimization time, indicating that the model is effectively utilizing 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 4 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. However, it is important to note that since the RH method selects intervals one scenario at a time, the iteration time does not depend on the number of scenarios, suggesting that this method could perform well in cases with a large number of scenarios.

Fig. 3 Optimization time over iterations.

Fig. 4 Iteration time over iterations.

5 Conclusion and Future Directions

The examples above demonstrate that aggregating time steps, combined with iterative refinement, can effectively solve the Capacity Expansion Problem (CEP).

We explored two different approaches for selecting time intervals to refine at each iteration. The first approach employed a validation function based on the Rolling Horizon method, 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 ρ, as defined in 3, 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 ρ across each node in the network served as a quality index for each time interval, guiding the disaggregation process by targeting intervals with the worst ρ values. We also provided a theoretical justification for using ρ, 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 a feasible solution for the original problem.

The iteration method with RH validation proved to be very effective at selecting relevant intervals to disaggregate. Starting with a day-night aggregation consisting of 730 intervals throughout the year, the method halted on average after XXX iterations with an RH-feasible solution whose cost diverged only minimally from the optimum. However, the use of RH within the iteration loop has a non negligible computational cost, granting it does not increase with the number of scenarios. Still, there is a lot of wriggle space in the implementation of this method, which can help further reduce the computational load.

The iteration method with ρ proved more effective than chance at selecting relevant intervals to disaggregate, while maintaining similar computational times. In future work, we plan to relax the conditions in Proposition 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 ρ, 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.

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, URL https://www.sciencedirect.com/science/article/pii/S0378779620301553

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, URL https://www.sciencedirect.com/science/article/pii/S1364032117311310

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, URL https://www.sciencedirect.com/science/article/pii/S0360319919345926

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, URL https://www.sciencedirect.com/science/article/pii/S037877881100329X

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), statistical Reports. Data is published based on aggregations of Transparency Platform data, complying with the Transparency regulation as described in the Detailed Data Description document.

ENTSO-E Statistical Reports (2024) Entso-e power statistics and transparency platform. URL https://www.entsoe.eu/data/power-stats/, statistical Reports. Data is published based on aggregations of Transparency Platform data, complying with the Transparency regulation as described in the Detailed Data Description document.

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

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, URL https://www.sciencedirect.com/science/article/pii/S0377221721006536

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, URL https://www.sciencedirect.com/science/article/pii/S0360544210000216

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, URL https://www.sciencedirect.com/science/article/pii/S019689041300589X

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, URL https://www.sciencedirect.com/science/article/pii/S1876610217330102

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, URL https://www.sciencedirect.com/science/article/pii/S0960148109001001

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, URL https://www.sciencedirect.com/science/article/pii/S1364032118307421

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

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, URL https://www.sciencedirect.com/science/article/pii/S1568494619303308

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 project, 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:

$$f(x;\theta,\gamma) = \left(\frac{\gamma}{\theta}\right) x^{\gamma-1} \exp\left(-\left(\frac{x}{\theta}\right)^{\gamma}\right) \tag{15}$$

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

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

$$\log L(\theta,\gamma) = \sum_{i=1}^{n} \log f(X_i \mid \theta,\gamma) \tag{16}$$

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

$$\frac{\partial \log L}{\partial \theta} = -\frac{n\gamma}{\theta} + \frac{\gamma}{\theta^2} \sum_{i=1}^{n} x_i^{\gamma} = 0 \tag{17}$$

Eliminating θ, we get:

$$\left[\frac{\sum_{i=1}^{n} x_i^{\gamma} \log x_i}{\sum_{i=1}^{n} x_i^{\gamma}} - \frac{1}{\gamma}\right] = \frac{1}{n} \sum_{i=1}^{n} \log x_i \tag{18}$$

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:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\gamma}}\right)^{\frac{1}{\hat{\gamma}}} \tag{19}$$

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:

$$\begin{cases} \alpha = \mathbb{E}[X]\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \\ \beta = (1-\mathbb{E}[X])\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \end{cases} \tag{20}$$

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

$$C(F_{Y_1}(y_1), \dots, F_{Y_T}(y_{|T|})) = P(Y_1 \le y_1, \dots, Y_{|T|} \le y_{|T|}). \tag{21}$$

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 $\{y_{t,j}\}_{t\in t, j\in J}$ 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 $\{\Phi^{-1}(\hat{F}_{Y_t}(y_{t,j}))\}_{t\in T, j\in J}$, 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 STRUCTURE PRESERVING CONTRAINT TRASFORMATIONS

Varying time aggregation can be viewed as performing row and column aggregation on the original linear programming (LP) model. Consider the following general linear problem:

$$\min_{x\in\mathbb{R}^n} c^T x \tag{22}$$

$$\text{s.t.} \quad Ax = b \tag{23}$$

$$x \ge 0 \tag{24}$$

Here, A is an $m \times n$ matrix. Now, let $\sigma = \{S_1, S_2, \ldots, S_{\tilde{n}}\}$ be a partition of $[n]$ (the columns) and $\delta = \{R_1, R_2, \ldots, R_{\tilde{m}}\}$ a partition of $[m]$ (the rows), corresponding to a partition of the rows and columns of A.

We obtain the corresponding aggregated problem by replacing each set S in σ with a single row, and each set R in δ with a single column. One way to aggregate a set of rows (or columns) is by taking a linear combination of the rows (or columns), known as *weighted aggregation*. We denote the weights of the aggregation by ω_r for $r \in \sigma$, and τ_c for $c \in \delta$. The corresponding aggregated LP problem becomes:

$$\min_{\tilde{x} \in \mathbb{R}^{\tilde{n}}} \tilde{c}^T \tilde{x} \tag{25}$$

$$\text{s.t.} \quad \tilde{A}\tilde{x} = \tilde{b} \tag{26}$$

$$\tilde{x} \geq 0 \tag{27}$$

where $\tilde{A}$ is a $\tilde{m} \times \tilde{n}$ matrix.

In the problem under consideration, we have various types of constraints: Electricity Balance, Hydrogen Balance, Hydrogen Storage, and bounds on the variables. Given a time partition P, we define σ and δ such that each set $S \in \sigma$ corresponds to all constraints of the same type, scenario, and time index t that falls within the same time interval in T as P. Similarly, the variables (such as Power generation, Hydrogen generation, etc.) are partitioned in δ based on the same criteria. Rows and columns are combined via weight aggregation. This aggregation maintains the structure of the original problem, meaning that had we formulated the model directly with the aggregated time steps, we would have arrived at the same model. Before defining a *structure-preserving aggregation* for a general LP, we introduce some notation: Given a matrix B with row and column index sets I and J, respectively, for any subsets $I' \subset I$ and $J' \subset J$, we denote the submatrix of B with rows in I' and columns in J' as $B_{I',J'}$.

Let $\tilde{A}$ be formed by aggregating the rows and columns of A according to the partitions σ and δ, respectively. For each $R \in \sigma$, we denote by $\tilde{A}_R$ the row in $\tilde{A}$ resulting from aggregating the rows of A corresponding to R, while A_R refers to the submatrix of A consisting of all rows in R. Similarly, for each $C \in \delta$, we define $\tilde{A}_C$ as the column in $\tilde{A}$ obtained by aggregating the columns of A in C, and A_C as the submatrix of A containing all columns in C. Thus, σ and δ serve as the index sets for $\tilde{A}$.

For a family of sets F, we denote the subsets of F with size exactly k and greater than k by $F_{=k}$ and $F_{>k}$, respectively. Specifically, $\text{supp}(\tilde{A}_R)_{>1} \subset \delta$ represents the set of indices corresponding to partitions $C \in \delta$ with size greater than 1, and $\text{supp}(A_r)_{>1}$ refers to the set of indices where $c \in C \in \delta$, with C having a size greater than 1.

Definition 4 Given an LP problem (22), we say that a weighted aggregation with respect to partitions σ, δ is *structure-preserving* if for each $R \in \sigma$ and each $r \in R$, there exists $f^r : [\tilde{n}] \to [n]$ such that:

1. $f^r|_{\text{supp}(\tilde{A}_R)} : \text{supp}(\tilde{A}_R)_{>1} \to \text{supp}(A_r)_{>1}$ is a bijection such that

$$\tilde{A}_{R,C} = A_{r,f^r(C)} \text{ for all } C \in \text{supp}(\tilde{A}_R)_{>1}$$

.

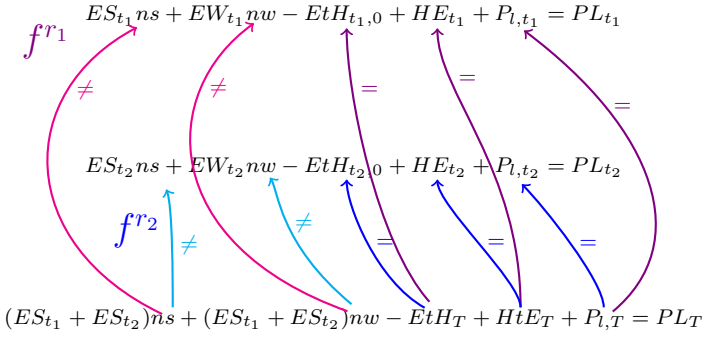
2. If $f^{r'}(C') = f^r(C)$ then $C = C'$.
3. $C = \{f^r(C)\}_{r \in R}$ for all $R \in \sigma_{>1}$ and $C \in \delta$.

From condition 3 follows the following:

Observation B.1 *For all $\{c\} \in \delta_{=1}$ and constraints $r \in R \in \sigma_{>1}$, $f^r(\{c\}) = c$.*

Example 1 As an example, we consider the function f^r corresponding to the time step aggregation in CEP as defined in subsection 3.1. Let us examine the Power Balance constraints r_1 and r_2 for a fixed node $n \in \mathcal{N}$ and time steps t_1 and t_2, where $T \coloneqq \{t_1, t_2\} \in P$ represents an interval in the time partition of the aggregated problem. We note that both f^{r_1} and f^{r_2} satisfy the conditions outlined in Definition 4.

Condition **??** is met as all aggregated variables, EtH_T, HtE_T, and $P_{l,T}$, are mapped (through violet arrows for f^{r_1} and blue arrows for f^{r_2}) to unaggregated variables with the same coefficients. Condition 3 implies that all variables appearing in the two constraints are included in the image of either f^{r_1} or f^{r_2}. Finally, Condition **??** holds trivially.



This implies that the coefficients of the aggregated variables in the aggregated problem match those in the original problem for the corresponding unaggregated variables, f^r can be seen as a function mapping the aggregated variables to variables of the same "type" in the unaggregated constraint. While obtaining a feasible solution to (22) from (25) is not always guaranteed, it is possible under certain assumptions. .

Observation B.2 *If (σ, δ) is a structure-preserving aggregation, let $R \in \sigma$ and $r \in R$. Let $\tilde{x}$ be a solution to the aggregated problem (25). If $\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}} \neq 0$, define*

$$\rho_r := \frac{b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}}}{\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}}.$$

If $A_{r,\delta_{=1}} = 0$ and $b_r = 0$ for all $r \in R$, then ρ_r can be chosen arbitrarily.

If $\rho_r \geq 0$ and $x \in \mathbb{R}^n$ satisfies $x_{\delta_{=1}} = \tilde{x}_{\delta_{=1}}$ and $x_{f^r(C)} = \rho_r \tilde{x}_C$ for all $C \in \text{supp}(\tilde{A}_R)_{>1}$, then x satisfies the constraint $A_r x = b_r$ of the original problem.

Proof Consider, $A_r x = \sum_{i \in \text{supp}(A_r)} A_{r,i} x_i$. This sum can be divided over the aggregated and unaggregated variables:

$$A_r x = A_{r,\delta_{=1}} x_{\delta_{=1}} + \sum_{c \in \text{supp}(A_r)_{>1}} A_{r,c} x_c. \tag{28}$$

If $A_{r,\delta_{=1}} = 0$ and $b_r = 0$ then $A_{r,\delta_{=1}} x_{\delta_{=1}} = 0$. Fix $\rho_r \geq 0$, then from the definition of structure-preserving aggregation, we know that $f^r(\text{supp}(\tilde{A}_R)_{>1}) = \text{supp}(A_r)_{>1}$, so equation (28) becomes:

$$(28) = \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} A_{R,f^r(S)} x_{f^r(S)} = \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \rho_r \tilde{x}_S = \rho_r \tilde{A}_R \tilde{x} = 0, \tag{29}$$

where the second equality holds because $\tilde{A}_{R,S} = A_{r,f^r(S)}$ and $x_{f^r(S)} = \rho_r \tilde{x}_S$. Thus, x satisfies the constraint $A_r x = b_r$.

When $A_{r,\delta_{=1}} \neq 0$ or $b_r \neq 0$, we proceed similarly:

$$(28) = A_{r,\delta_{=1}} x_{\delta_{=1}} + \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} A_{r,f^r(S)} x_{f^r(S)} \tag{30}$$

$$= A_{r,\delta_{=1}} \tilde{x}_{\delta_{=1}} + \rho_r \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \tilde{x}_S. \tag{31}$$

By the definition of ρ_r the first sum in the last line of equation (30) is equal to:

$$\begin{aligned}\rho_r \sum_{S\in \text{supp}(\tilde{A}_R)>1} \tilde{A}_{R,S}\tilde{x}_S &= \rho_r(\tilde{A}_R\tilde{x} - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}) \\ &= \rho_r(\tilde{b}_R - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}) \\ &= b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}}.\end{aligned}$$

Thus, we obtain:

$$A_r x = b_r.$$

□

A structure-preserving aggregation does not inherently ensure the feasibility of all constraints in the original problem. However, Observation B.2 demonstrates how to partially reconstruct a solution x for a specific constraint r by scaling the aggregated variables appropriately within the support of A_r.

Definition 5 Let ρ_r be defined as in Observation B.2 for all $r \in R \in \sigma_{>1}$. Let $x \in \mathbb{R}^n$ be defined as $x_{\delta_{=1}} \coloneqq \tilde{x}_{\delta_{=1}}$ and $x_{f^r(C)} \coloneqq \rho_r\tilde{x}_C$ for all $C \in \delta_{>1}$ and $r \in R \in \sigma_{>1}$. Then x is well defined if for all $r, r' \in R \in \sigma_{>1}$ such that $f^r(C) = f^{r'}(C)$, we have $\rho_r = \rho_{r'}$. In such case we refer to x as a *rho*-solution.

If x is a ρ-solution. x is a feasible solution for the constraints in $\sigma_{>1}$, we also need to ensure that x is also feasible for the remaining constraints in $\sigma_{=1}$.

Example 2 Consider the constraint that the initial hydrogen stored must be equal to the final hydrogen stor for a one node network:

$$\sum_{t=1}^{n} \Delta H_t = 0 \tag{32}$$

The corresponding aggregated constraint is for a time partition P is:

$$\sum_{T\in P} \Delta H_T = 0 \tag{33}$$

For all ρ such that $\sum_{t\in T} \rho_t = 1$ for all T in P, given a feasible solution for the aggregated problem $\Delta\tilde{H}_T$, let $\Delta H_t \coloneqq \rho_t\Delta\tilde{H}_T$, then constraint (32) holds:

$$\sum_{t=1}^{n} \Delta H_t = \sum_{t=1}^{n} \rho_t\Delta\tilde{H}_T = \sum_{T\in P}\sum_{t\in T} \rho_t\Delta\tilde{H}_T = \sum_{T\in P} \Delta\tilde{H}_T = 0 \tag{34}$$

Thus constraint (32)is holds for ρ-solutions

This is a special instance of a general class of constraints that always hold for ρ-solutions, this follows from the following propriety of ρ-solutions:

Observation B.3 *Let $\omega_r \in \mathbb{R}$ for all $r \in R \in \sigma$ be the weights of the row aggregation. If x is a ρ-solution, then we have, for all $R \in \sigma_{>1}$:*

$$\omega_R^T\rho_R = 1 \tag{35}$$

Proof

$$\omega_R^T\rho_R = \sum_{r\in R} \omega_r\rho_r = \frac{\sum_{r\in R} \omega_r(b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}})}{\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}} = 1$$

Note that if x is a ρ-solution, then for all $C \in \text{supp}(A_r)_{>1}$, we can pick $R^{(C)} \in \sigma_{>1}$ so that $C \in \text{supp}(\tilde{A}_{R^{(C)}})$ and $x_{f^r(C)} = \rho_r \tilde{x}_C$ for all $r \in R^{(C)}$ and the definition of x does not depend on the choice of $R^{(C)}$.

Observation B.4 *Le (σ, δ) be a structure-preserving, row and column aggregation. If x is a ρ-solution and r is a constraint in $\sigma_{=1}$, such that*

$$A_{r,f^{r'}(C)} = \omega_{r'} \tilde{A}_{r,C} \text{ for all } r' \in R^{(C)},\ C \in \delta_{>1},$$

then x is a fesible solution for constraint r.

Proof As before we split the sum $A_r x$ over aggregated and unaggregated variables:

$$A_r x = \sum_{C \in \delta_{=1}} A_{r,C} x_C + \sum_{C \in \text{supp}(A_r)_{>1}} \sum_{r' \in R^{(C)}} A_{r,f^{r'}(C)} x_{f^{r'}(C)} \tag{36}$$

From the hypothesis and the definition of ρ-solution, we have:

$$(36) = \sum_{C \in \delta_{=1}} \tilde{A}_{r,C} \tilde{x}_{r,C} + \sum_{C \in \text{supp}(A_r)_{>1}} \sum_{r' \in R^{(C)}} \omega_{r'} \tilde{A}_{r,C} \rho_{r'} \tilde{x}_C \tag{37}$$

Since $\omega_R \rho_R = 1$, we have

$$(37) = \sum_{C \in \delta_{=1}} \tilde{A}_{r,C} \tilde{x}_{r,C} + \sum_{C \in \text{supp}(A_r)_{>1}} \tilde{A}_{r,C} \tilde{x}_C = \tilde{A}_r \tilde{x}_r = \tilde{b}_r = b_r \tag{38}$$

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

Definition 6 The *hypergraph associated to the aggregation* (σ, δ) is the hypergraph $\mathcal{N}, \mathcal{E}$ having as nodes the aggregated variables $\mathcal{N} \coloneqq \delta_{>1}$ and as edges the subsest of $\mathcal{N}$ that appear together in aggregated constraints.

When two edges (constraints) in the hypergraph, r and r', share aggregated variables, the scaling factors ρ_r and $\rho_{r'}$ must be equal for Observation B.2 to hold for both r and r'. From this follows the following:

Proposition 3 *If (σ, δ) is a structure-preserving aggregation and the constraints in $\sigma_{=1}$ hold for rho-solutions. Let $\tilde{x}$ be a solution to the aggregated problem* (25)*. For all $r \in R \in \sigma_{>1}$ define ρ_r as in Observation B.2. If $\rho_r \geq 0$ and is constant over the connected components of the hypergraph associated to (σ, δ). Then $x_{\delta_{=1}} \coloneqq \tilde{x}_{\delta_{=1}}$ and $x_{f^r(C)} \coloneqq \rho_r \tilde{x}_C$ for all $C \in \text{supp}(\tilde{A}_R)$ and $C \in \delta_{>1}$ is well defined and thus x is a ρ-solution. Further more if x is feasible for all constraints in $\delta_{=1}$, then x is feasible solution to the unaggregated problem* (22)*.*

Until now we have only considered feasibility, ignoring the relationship between the cost of $\tilde{x}$ and the cost of x. The following observation gives a condition for the cost of $\tilde{x}$ to be equal to the cost of x. For all $C \in \delta_{>1}$ let $R^{(C)} \in \sigma_{>1}$ be so that $C \in \text{supp}(\tilde{A}_{R^{(C)}})$ and $x_{f^r(C)} = \rho_r \tilde{x}_C$ for all $r \in R^{(C)}$.

Observation B.5 *Let x be a ρ-solution. If $\omega_r \tilde{c}_C = c_{f(r,C)}$ for all $r \in R^{(C)} \in \sigma_{>1}$, then the cost of $\tilde{x}$ for the aggregated problem is equal to the cost of x in the unaggregated problem.*

Proof Let $\tilde{x}$ be a solution to the aggregated problem (25). Using observation B.3, for all $C \in \delta_{>1}$ the cost corresponding to the variable $\tilde{x}_C$ is

$$\tilde{c}_C \tilde{x}_C = \tilde{c}_C (\sum_{r \in R^{(C)}} \omega_r \rho_r) \tilde{x}_C = \sum_{r \in R^{(C)}} \tilde{c}_C \omega_r \rho_r \tilde{x}_C = \sum_{r \in R^{(C)}} c_{f(r,C)} x_{f(r,C)}$$

Which corresponds to the cost of the variables $\{x_{f(r,C)}\}_{r\in R^{(C)}}$. Thus

$$\begin{aligned}\tilde{c}\tilde{x} &= \sum_{C\in\delta_{=1}} \tilde{c}_C\tilde{x}_C + \sum_{C\in\delta_{>1}} \tilde{c}_C\tilde{x}_C \\ &= \sum_{C\in\delta_{=1}} c_C x_C + \sum_{C\in\delta_{>1}} \sum_{r\in R^{(C)}} c_{f(r,C)}x_{f(r,C)} \\ &= \sum_{C\in\delta_{=1}} c_C x_C + \sum_{j\in\cup_{C\in\delta_{>1}}C} c_j x_j \\ &= cx\end{aligned}$$

□

While row aggregation of a linear problem is a relaxation of the original problem, the same does not apply to column aggregation. However, 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 7 A column aggregation of a linear problem, respect to the column partition δ, is a *constant-coefficients column aggregation* if, for all $C \in \delta_{>1}$, the non-zero rows in A_C are identical.

Proposition 4 *If (σ, δ) is a structure-preserving, constant-coefficients column aggregation, if the hypothesis of proposition 3 and observation B.5 are true, then the aggregated problem* (25) *is exact and x is an optimal solution.*

Since for observation B.5, the cost of the aggregated problem is equal to the cost of x in the unaggregated problem, we only need 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 are relaxations.