Model Reduction in Capacity Expansion Planning Problems via Renewable Generation Site Selection

David Radu, Antoine Dubois, Mathias Berger, Damien Ernst

Dept. of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, Belgium
{dcradu, antoine.dubois, mathias.berger, dernst}@uliege.be

Abstract—The accurate representation of variable renewable generation (RES, e.g., wind, solar PV) assets in capacity expansion planning (CEP) studies is paramount to capture spatial and temporal correlations that may exist between sites and impact both power system design and operation. However, it typically has a high computational cost. This paper proposes a method to reduce the spatial dimension of CEP problems while preserving an accurate representation of renewable energy sources. A two-stage approach is proposed to this end. In the first stage, relevant sites are identified via a screening routine that discards the locations with little impact on system design. In the second stage, the subset of relevant RES sites previously identified is used in a CEP problem to determine the optimal configuration of the power system. The proposed method is tested on a realistic EU case study and its performance is benchmarked against a CEP set-up in which the entire set of candidate RES sites is available. The method shows great promise, with the screening stage consistently identifying 90% of the optimal RES sites while discarding up to 54% of the total number of candidate locations. This leads to a peak memory reduction of up to 41% and solver runtime gains between 31% and 46%, depending on the weather year considered.

Index Terms—variable renewable generation, capacity expansion planning, spatial reduction, two-stage method

I. INTRODUCTION

Capacity expansion planning (CEP) problems are powerful tools for the design, analysis and implementation of energy system decarbonisation policies. In such frameworks, the accurate spatiotemporal representation of variable renewable energy generation (RES, e.g., wind, solar PV) is paramount for the precise estimation of capacity requirements [1]. However, the detailed modelling of RES comes at a high computational cost and ways to mitigate this issue in order to strike the right balance between accuracy and computational effort when solving such problems are necessary, yet seldom proposed. For example, a highly detailed representation of RES within a CEP set-up cast as a linear program (LP) is proposed by MacDonald et al. [2], yet the reported runtimes (thousands of core hours for large-scale instances) limit its use in practice and its reproducibility. Wu et al. [3] also propose an LP-cast CEP framework in which high-resolution RES modelling is made possible via a GIS-based resource assessment tool. Nonetheless, the coefficient matrix stores hourly capacity factor values at each location and is therefore full, which limits the scalability of the proposed method to a few hundreds of candidate RES sites only, thus rendering it unsuitable for large-scale applications.

Although plenty of work has been carried out in recent years to develop temporal reduction techniques for RES in CEP settings [4], studies tackling the issue of spatial model reduction are scarce. Cohen et al. [5] suggest the aggregation of RES in resource regions, with wind and solar PV resources over the contiguous United States being modelled via 356 and 134 profiles, respectively. In a similar vein, Hörsch and Brown [6] leverage a CEP framework formulated as an LP to assess the impact of spatial resolution on the outcomes of co-optimizing generation and transmission assets across Europe. A network reduction process based on k-means clustering is incorporated in their method and the resulting topology serves as the basis for modelling renewable resources. More precisely, Europe-wide RES are represented via 37 to 362 different aggregate profiles, depending on the desired number of network clusters. While spatial aggregation approaches, as the ones proposed in [5], [6], partly mitigate the aforementioned computational issues [2], [3], the limited number of RES profiles considered hinders their ability to exploit the benefits of resource diversity which, in turn, can lead to system cost overestimation [7].

This paper proposes a method to reduce the spatial dimension and decrease the computational requirements of CEP problems while preserving a detailed representation of RES assets. This is achieved by leveraging a two-stage heuristic that can be described as follows. The first stage, which is cast as an LP, is used to screen a set of candidate sites and identify sites that have little impact on optimal system design, which are then discarded. In the second stage, information (geo-positioning and capacity factors time series) about the remaining sites is used as input data in a CEP framework that determines the installed capacities of generation, storage and transmission assets leading to a minimum-cost system configuration. Thus, the proposed method makes it possible to reduce the size of the CEP problem, and therefore enables memory and computation time savings.

The paper is structured as follows. Section II details the methods at the core of the proposed two-stage approach. Then, Section III briefly describes the case study used to showcase the applicability of the suggested approach before results are reported in Section IV. Section V concludes the paper and discusses future work avenues.

II. METHOD

The proposed solution method (or SM) is introduced in this section. Firstly, the standard CEP framework (from hereon, the FLP) is formulated. In the remainder of this paper, the FLP denotes the CEP set-up that simultaneously tackles the siting and sizing of RES assets, as well as the sizing of other power system (e.g., generation, storage or transmission) technologies. Then, the screening method for candidate RES sites (SITE) that enables the formulation of a reduced-size CEP framework (from hereon, the RLP) is described. The SITE-RLP sequence will hereafter be referred to as the SM.

A. Capacity expansion planning framework

Let $\mathcal{N}_B$ and $\mathcal{L}$ be the sets of existing buses and transmission corridors, respectively. Let $\mathcal{N}_R$ be a set of candidate RES sites that may be connected to buses $n \in \mathcal{N}_B$, which is partitioned into disjoint subsets $\mathcal{N}_R^n$. The CEP formulation reads

$$\min_{\mathbf{K}, (\mathbf{p}_t)_{t \in \mathcal{T}}} \omega \Big[\sum_{\substack{n \in \mathcal{N}_B \\ m \in \mathcal{N}_R^n}} \left(\zeta^m + \theta_f^m \right) K_{nm} + \sum_{\substack{n \in \mathcal{N}_B \\ j \in \mathcal{G} \cup \mathcal{S}}} \left(\zeta^j + \theta_f^j \right) K_{nj}$$

$$+ \sum_{l \in \mathcal{L}} \left(\zeta^l + \theta_f^l \right) K_l \Big] + \sum_{t \in \mathcal{T}} \Big[\sum_{l \in \mathcal{L}} \theta_v^l |p_{lt}| \tag{1a}$$

$$+ \sum_{\substack{n \in \mathcal{N}_B \\ m \in \mathcal{N}_R^n}} \theta_v^m p_{nmt} + \sum_{\substack{n \in \mathcal{N}_B \\ j \in \mathcal{G} \cup \mathcal{S}}} \theta_v^j |p_{njt}| + \sum_{n \in \mathcal{N}_B} \theta^e p_{nt}^e \Big]$$

s.t. $\displaystyle \sum_{m \in \mathcal{N}_R^n} p_{nmt} + \sum_{g \in \mathcal{G}} p_{ngt} + \sum_{s \in \mathcal{S}} p_{nst}^D + \sum_{l \in \mathcal{L}_n^+} p_{lt} + p_{nt}^e$

$$= \lambda_{nt} + \sum_{s \in \mathcal{S}} p_{nst}^C + \sum_{l \in \mathcal{L}_n^-} p_{lt}, \ \forall n \in \mathcal{N}_B, \forall t \in \mathcal{T} \tag{1b}$$

$$p_{nmt} \leq \pi_{nmt}(\kappa_{nm}^0 + K_{nm}),$$
$$\forall n \in \mathcal{N}_B, \forall m \in \mathcal{N}_R^n, \forall t \in \mathcal{T} \tag{1c}$$

$$\kappa_{nm}^0 + K_{nm} \leq \bar{\kappa}_{nm}, \ \forall n \in \mathcal{N}_B, \forall m \in \mathcal{N}_R^n \tag{1d}$$

$$p_{ngt} \leq \kappa_{ng}^0 + K_{ng}, \ \forall n \in \mathcal{N}_B, \forall g \in \mathcal{G}, \forall t \in \mathcal{T} \tag{1e}$$

$$\kappa_{ng}^0 + K_{ng} \leq \bar{\kappa}_{ng}, \ \forall n \in \mathcal{N}_B, \forall g \in \mathcal{G} \tag{1f}$$

$$p_{nst} = -p_{nst}^C + p_{nst}^D, \forall n \in \mathcal{N}_B, \forall s \in \mathcal{S}, \forall t \in \mathcal{T} \tag{1g}$$

$$|p_{nst}| \leq \phi_s(\kappa_{ns}^0 + K_{ns}), \forall n \in \mathcal{N}_B, \forall s \in \mathcal{S}, \forall t \in \mathcal{T} \tag{1h}$$

$$e_{nst} \leq \kappa_{ns}^0 + K_{ns}, \forall n \in \mathcal{N}_B, \forall s \in \mathcal{S}, \forall t \in \mathcal{T} \tag{1i}$$

$$e_{nst} = \eta_s^{SD} e_{ns(t-1)} + \eta_s^C p_{nst}^C - p_{nst}^D/\eta_s^D,$$
$$\forall n \in \mathcal{N}_B, \forall s \in \mathcal{S}, \forall t \in \mathcal{T} \tag{1j}$$

$$\kappa_{ns}^0 + K_{ns} \leq \bar{\kappa}_{ns}, \ \forall n \in \mathcal{N}_B, \forall s \in \mathcal{S} \tag{1k}$$

$$|p_{lt}| \leq \kappa_l^0 + K_l, \ \forall l \in \mathcal{L}, \forall t \in \mathcal{T} \tag{1l}$$

$$\kappa_l^0 + K_l \leq \bar{\kappa}_l, \ \forall l \in \mathcal{L} \tag{1m}$$

The problem described in (1a-m) minimizes total system cost subject to a set of constraints of the underlying assets. The objective function (1a) comprises capital expenditure, fixed and variable operating costs of the generation, storage and transmission assets, as well as the economic penalties associated with unserved demand. Constraint (1b) enforces the energy balance at each bus, while the operation and sizing of

RES assets is modelled via (1c-d). Note that a single RES technology $r \in \mathcal{R}$ is associated with each site $m \in \mathcal{N}_R$. Then, conventional generators are modelled via (1e-f) and the operation and sizing of storage units follows (1g-k). Finally, constraints (1l-m) encode the transportation model governing the power flows in transmission links. It is worth noting that, although the absolute values in Eqs. (1a), (1h) or (1l) render the CEP problem described in (1a-m) non-linear, it can be cast as an LP using standard reformulation techniques.

B. Renewable sites selection method

The proposed SM works by decoupling the siting and sizing of RES assets. At first, the SITE stage is leveraged to screen the sets of candidate RES locations and identify those sites that play a role in the optimal system design, while discarding the rest. To this end, the siting problem is formulated by i) discarding some complicating variables and approximating a subset of complicating constraints (i.e., the ones associated with dispatchable power generation, storage systems and power flows in transmission lines) and ii) relaxing and taking linear combinations, as well as scaling the right-hand site coefficients of certain equality constraints (i.e., the power balance equations). The objective function (2a) is obtained by preserving the terms related to the costs of deploying and operating RES technologies and the economic penalty associated with unserved demand. Then, the constraints discarded from (1a-m) are approximated via two parameters found in (2b). More formally, let $\mathcal{T}$ be the set of time periods, let $\mathcal{T}_\tau \subseteq \mathcal{T}$, $|\mathcal{T}_\tau| = \delta\tau$, $\tau = 1, \ldots, T$, be a collection of disjoint subsets forming a partition of $\mathcal{T}$ into time slices of length $\delta\tau$. More precisely, $\delta\tau$ represents the length of a time slice (e.g., one hour, one day) over which the energy balance in (2b) is enforced and its role is to emulate the behavior of storage assets shifting RES supply in time. Furthermore, let $\xi_\tau^n \in \mathbb{R}_+$ denote regional minimum RES feed-in targets enforced over every time slice $\mathcal{T}_\tau$, $\tau = 1, \ldots, T$. This parameter enforces a minimum level of local power production from renewable sources which i) mirrors the effect of transmission constraints and ii) accounts for low-carbon legacy generation capacity that would offset the country-specific RES requirements. Constraints (1c-d) are preserved as such and the siting problem thus reads

$$\min_{\mathbf{K}, (\mathbf{p}_t)_{t \in \mathcal{T}}} \omega \Big[\sum_{\substack{n \in \mathcal{N}_B \\ m \in \mathcal{N}_R^n}} \left(\zeta^m + \theta_f^m \right) K_{nm} \Big] +$$

$$\sum_{t \in \mathcal{T}} \Big[\sum_{\substack{n \in \mathcal{N}_B \\ m \in \mathcal{N}_R^n}} \theta_v^r p_{nmt} + \sum_{n \in \mathcal{N}_B} \theta^e p_{nt}^e \Big] \tag{2a}$$

s.t. $\displaystyle \sum_{t \in \mathcal{T}_\tau} \Big[\sum_{m \in \mathcal{N}_R^n} p_{nmt} + p_{nt}^e \Big] \geq \xi_\tau^n \sum_{t \in \mathcal{T}_\tau} \lambda_{nt},$

$$\forall n \in \mathcal{N}_B, \ \forall \tau \in \{1, \ldots, T\} \tag{2b}$$

$$p_{nmt} \leq \pi_{nmt}(\kappa_{nm}^0 + K_{nm}),$$
$$\forall n \in \mathcal{N}_B, \forall m \in \mathcal{N}_R^n, \forall t \in \mathcal{T} \tag{2c}$$

$$\kappa_{nm}^0 + K_{nm} \leq \bar{\kappa}_{nm}, \ \forall n \in \mathcal{N}_B, \forall m \in \mathcal{N}_R^m \tag{2d}$$

For every $n \in \mathcal{N}_B$, the problem returns the set of candidate RES sites identified as relevant (with an installed capacity above $1 \, \mathrm{MW}$) in the optimal system design, i.e. $\mathcal{N}_{\mathrm{SITE}}^n$. Then, the RLP is built by replacing $\mathcal{N}_R^n$ with $\mathcal{N}_{\mathrm{SITE}}^n$ in constraints (1a-d) of the CEP problem.

III. CASE STUDY

Input Data: The analysis is conducted for three individual weather years (i.e., 2016, '17 and '18) and over 33 countries within the *ENTSO-E* system. The siting stage relies on hourly-sampled resource data obtained from the *ERA5* reanalysis database [8] at a spatial resolution of $1.0°$. The mapping of resource data to capacity factors time series is achieved via the transfer functions of appropriate conversion equipment for each individual technology. More precisely, a site-specific selection of wind generators is carried out based on the *IEC 61400* standard [9] and four different converters are available for deployment (i.e., the *Vestas V110, V90, V117* and *V164*), each of them suitable for specific wind regimes. The selection of solar energy converters is done on a technology basis, with the *TrinaSolar DEG15MC* module available for utility-scale PV deployment and the *TrinaSolar DD06M* array available for distributed PV generation. A greenfield approach is adopted, i.e., no legacy capacity of RES assets is considered, while the technical potential is estimated via a land eligibility assessment framework [10] that yields eligible surface areas for RES deployment for a set of 1740 candidate sites. A set of assumptions pertaining to the power densities of different generation technologies are then made to map surface areas into maximum allowable installed capacities, i.e., technical potentials. Specifically, a density of $5 \, \mathrm{MW/km^2}$ is considered for wind deployments [11]. With respect to solar PV units, power densities of $40 \, \mathrm{MW/km^2}$ and $16 \, \mathrm{MW/km^2}$ are considered for utility-scale and residential installations, respectively [12]. Electricity demand time series for all considered countries are retrieved from the *OPSD* platform [13].

The CEP frameworks (i.e., both the FLP and RLP) follow a centralized planning approach and build upon the 2018 TYNDP dataset, where each European country is modelled as one node [14]. The resulting network topology is displayed in Fig. 1. In this exercise, the expansion of the transmission network is limited to the reinforcement of existing links. Furthermore, the total capacity of each link may not exceed twice the 2040 capacity estimated for this link in the TYNDP. Besides the four RES technologies sited in the previous stage, three more generation technologies are available for power generation, namely run-of-river (ROR) and reservoir-based (STO) hydro, as well as combined-cycle gas turbines (CCGT), with the latter being the only of the three that is also sized in (1a-m). The existing capacities of the other two are retrieved from [15], where the existence of $34 \, \mathrm{GW}$ of ROR and $98 \, \mathrm{GW}$ of STO installations is reported. Then, two technologies are available for electricity storage, namely pumped-hydro (PHS) units and Li-Ion batteries. The latter is the only one being sized in (1a-m) and a fixed energy-to-power ratio of $4 \, \mathrm{h}$ is assumed. The legacy capacity of the former is retrieved from
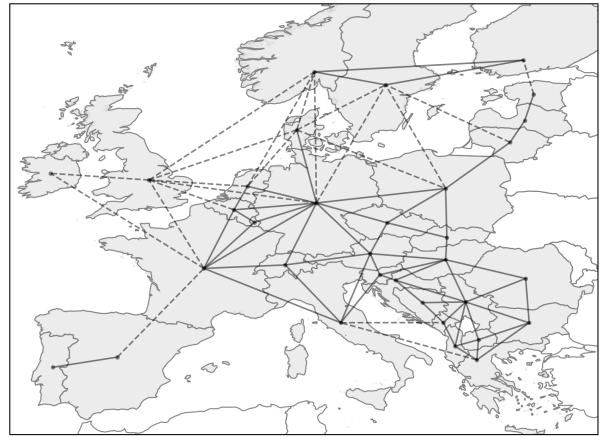

Fig. 1: System topology in the capacity expansion planning framework. AC connections displayed in full lines, DC links shown in dotted lines. The 33 nodes form the set $\mathcal{N}_B$ in (1a-m) and (2a-d).

[15], where $55 \, \mathrm{GW}/1950 \, \mathrm{GWh}$ of PHS storage is reported. The CEP problem is implemented in PyPSA 0.17 [16], while the techno-economic assumptions are gathered in [17].

Parametrization of the SITE stage: The two parameters of (2a-d) are defined as follows. First, the slicing period $\delta\tau$ is considered to be equal to $24 \, \mathrm{h}$, which corresponds to the nonzero frequency component of the aggregate EU-wide RES capacity factor time series with the largest amplitude (i.e., as provided by a discrete Fourier transform). Then, the country-dependent ξ_τ^n values are assumed not to be time-dependent and their estimation proceeds as follows. First, the residual demand (i.e., the difference between demand and generation potential of legacy dispachable units) is computed at peak load conditions. Then, the RES generation potential during the same time instants is determined. For each country, if RES potential exceeds the electricity demand for at least half the time steps in the optimization horizon, its potential transmission capabilities (i.e., 2040 TYNDP capacity limits times the length of slicing period $\delta\tau$) are added to the residual demand, as the country is a potential exporter of electricity in the EU-wide system. Conversely, if the electricity demand is higher than the RES potential most of the time, the transmission capabilities of that country are subtracted from the residual demand, as cross-border exchanges will oftentimes be used to cover for the domestic electricity needs. Finally, the ξ_τ^n values are determined as the ratio between the RES potential and the transmission capacity-adjusted residual demand, respectively.

Implementation: The SM, as well as the FLP are implemented in Python 3.7 and the proposed instances are run on a workstation running under CentOS, with an 18-core Intel Xeon Gold 6140 CPU clocking at $2.3 \, \mathrm{GHz}$ and $256 \, \mathrm{GB}$ RAM. Gurobi 9.0 was used to solve both (1a-m) and (2a-d). The dataset and code used in these simulations are available at [17] and [18].

IV. Results

The results of a set of experiments evaluating the performance of the SM against the FLP are detailed in this section.

Table I summarizes the performance of the siting stage by means of two indicators. First, the technology-specific spatial reduction share (γ_r) denotes the proportion of initial candidate RES sites discarded via SITE. Then, the screening accuracy (α_r) measures the ability of the method to identify the relevant candidate RES sites. More formally, let $\mathcal{R}$ be the set of renewable technologies and let $\mathcal{N}_R^r$ be the subset of sites with technology $r \in \mathcal{R}$ (these subsets are disjoint for different r and form a partition of $\mathcal{N}_R$). Note that for the purpose of this paper, offshore and onshore wind are considered as different resources. In addition, let $\mathcal{N}_{\mathrm{FLP}}^r$ and $\mathcal{N}_{\mathrm{SITE}}^r$ be the subsets of $\mathcal{N}_R^r$ selected by FLP and SITE where at least $1\,\mathrm{MW}$ of capacity is deployed, respectively. Then, the screening accuracy is defined as

$$\alpha_r = \frac{|\mathcal{N}_{\mathrm{SITE}}^r \cap \mathcal{N}_{\mathrm{FLP}}^r|}{|\mathcal{N}_{\mathrm{FLP}}^r|}, \ \forall r \in \mathcal{R}, \tag{3}$$

where $|\mathcal{N}|$ denotes the cardinality of set $\mathcal{N}$. First, in this table, it can be seen that the relative reduction achieved by SITE varies from 6% for utility-scale PV to 62% for distributed PV installations in the 2017 instance, with an average reduction in onshore and offshore wind sites of 38% and 54%, respectively. Furthermore, an overall reduction of the number of selected RES sites of up to 54% is observed across the three considered instances. In other words, less than half of the candidate RES sites are found to be relevant in the optimal system configuration by SITE and subsequently passed to the RLP. With respect to the ability of SITE to identify relevant RES locations, only the distributed PV sites have a selection accuracy score below 85%. However, the limited deployment of this technology in the solution of the proposed CEP instances enables the screening stage to properly identify over 90% of the relevant RES sites (i.e., the ones appearing in the FLP solution), irrespective of the weather year considered.

However, not all candidate RES sites found in the FLP solution are identified by SITE which selects different locations instead. For instance, when the latter is run with 2016 weather data, it fails to identify a total of 45 sites (14 onshore wind, 12 offshore wind and 19 distributed PV locations, respectively) out of 418 identified in the benchmark. Investigating how far these locations are from the ones selected by the FLP provides a first insight into how different the system designs associated

TABLE I: Technology-specific sites reduction (γ_r) and screening accuracy (α_r) of SITE. Number of candidate sites used in the FLP specified in parantheses.

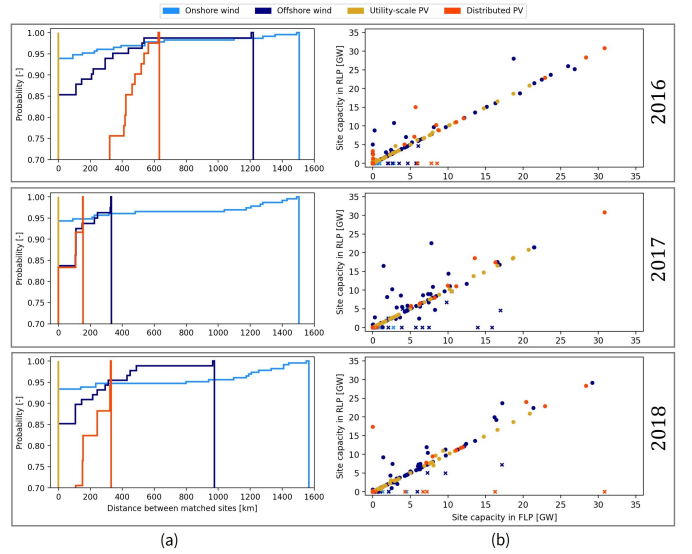|	$\mathrm{W_{on}}$ (590)		$\mathrm{W_{off}}$ (417)		$\mathrm{PV_u}$ (128)		$\mathrm{PV_d}$ (605)	
	γ_r	α_r	γ_r	α_r	γ_r	α_r	γ_r	α_r
2016	0.40	0.94	0.55	0.85	0.10	1.00	0.57	0.54
2017	0.37	0.94	0.55	0.86	0.06	1.00	0.62	0.83
2018	0.36	0.93	0.52	0.85	0.16	1.00	0.59	0.59

Fig. 2: (a) Distribution of geographical distances between pairs of sites identified via SITE and the FLP. (b) Site-specific installed capacity correlation between the RLP and the FLP.

with the two methods are. If the distances between the locations selected via SITE and FLP were found to be small, one would expect the effect of misidentifying sites to be limited, as RES patterns are usually comparable at neighboring sites. Conversely, large distances between sites identified via the two methods would often imply distinct RES patterns and could thus lead to substantial differences in the way the technologies are sized. The result of this analysis is shown in Fig. 2a. These plots depict, for each technology and weather year, the distribution of distances (expressed in kilometres) between pairs of sites selected via the FLP and SITE, respectively. The procedure used to generate these curves is as follows. First, distances of zero are associated to the pairs of sites found by both methods (α_r shares in Table I). Then, each unidentified site in the FLP solution is matched with the geographically closest (based on the geodesic distance) location in the set of SITE-exclusive locations. Once two sites are paired, none of them can be subsequently matched with another. Upon pairing all unidentified sites in the FLP with a counterpart in SITE, a cumulative distribution function of technology-specific distances is plotted. It can be observed in these three plots that, without exception, the 95th percentile of the matching distance for any of the four RES technologies falls below $500\,\mathrm{km}$. In a European context, it has been previously shown that country-aggregated wind output (usually more spatially heterogeneous than PV generation) is remarkably correlated at distances below the aforementioned threshold, especially in the North Sea basin where most onshore and offshore sites are deployed in the studied instances [19]. Furthermore, a maximum distance between matched sites of under $1600\,\mathrm{km}$ is reported for all technologies and weather years, with the largest discrepancies being consistently observed for onshore wind locations.

TABLE II: Differences in system-wide capacities between `FLP` and `RLP` for various technologies and weather years. A positive value reflects more capacity installed (or higher TSCE) in the `RLP`, while a negative value indicates more capacity in the `FLP`.

Year	W_{on} [GW]	W_{off} [GW]	PV_u [GW]	PV_d [GW]	CCGT [GW]	AC [TWkm]	DC [TWkm]	Li-Ion [GWh]	TSCE [b€]
2016	-7.38 (-16.31%)	6.06 (1.48%)	3.86 (1.92%)	-1.63 (-1.01%)	0.10 (0.11%)	-0.03 (-0.04%)	-0.05 (0.10%)	0.51 (1.99%)	0.36 (0.40%)
2017	-7.94 (-12.84%)	4.96 (1.17%)	0.60 (0.29%)	7.51 (7.36%)	1.01 (2.30%)	-2.33 (-2.90%)	-0.16 (-0.29%)	1.18 (4.07%)	0.44 (0.52%)
2018	-13.73 (-23.34%)	12.19 (2.97%)	2.51 (1.21%)	-5.60 (-2.99%)	1.03 (2.54%)	0.57 (0.65%)	-0.04 (-0.07%)	-1.53 (-3.63%)	0.42 (0.48%)

Upon screening the candidate RES locations via `SITE`, the `RLP` is run in order to retrieve, among others, the associated installed capacities. Fig. 2b shows, for each weather year, the correlation between installed capacities of i) the sites identified in the `FLP` and ii) the sites identified by `SITE` and sized via `RLP`. In this plot, round markers (o) denote data points associated with locations that are common to `FLP` and `SITE`, while crosses (x) represent data points corresponding to the pairs of sites matched according to the procedure described in the previous paragraph. The first remark in these plots is that in 76% (for 2016) to 79% (for 2018) of the cases, the installed capacities of `FLP` and `RLP` sites are matched to MW-order precision. Then, it can be observed that most of the (x) markers are situated on the bottom of the corresponding subplots. A complementary analysis of the resource signals associated with these data points suggests the existence of high-quality RES sites exploited by the `FLP`, but whose `SITE` counterparts (determined via the distance-based pairing algorithm) exhibit inferior resource quality and thus end up not being part of the `RLP` solution. In such a situation, the missing capacity, i.e., `FLP` capacity of the (x) data points in the lower part of the plot, is compensated in the `RLP` by superior power ratings at (o) sites above the trend line in Fig. 2b.

Table II reports, for different data years and for various technologies sized within the CEP stage, the difference between the system-wide installed capacities obtained by the `FLP` and `RLP` models, respectively (positive values indicate more capacity in the latter). In the last column, it can be seen that the relative objective function difference (i.e., the TSCE) between the two CEP set-ups does not exceed 0.52%, irrespective of the weather year considered. However, as suggested in a recent study by Neumann and Brown [20], rather small differences in total system costs can translate into fairly distinct system configurations. In this exercise, differences of 23.3%, 2.9%, 1.9% and 7.3% are reported for onshore wind, offshore wind,

utility-scale and distributed PV, respectively, between the `RLP` and the `FLP`. A closer look at the breakdown of capacities per country reveals the reasons behind such differences, as the large majority of the discrepancies observed in Table II are associated to a handful of resource-rich countries (e.g., Ireland, Italy, Spain or the UK). For instance, in 2017 and 2018, the `FLP` over-sizes onshore wind (and, thus, selects more sites) in Ireland and the UK, and uses it to supply Central Europe. Under the proposed $(\delta\tau, \xi_\tau^n)$ set-up of the `SITE` stage, a subset of these locations are not identified (see discussion on the (x) markers in Fig. 2a) and the associated capacity in the `FLP` is replaced in the `RLP` by a mix of offshore wind and distributed PV. Further on in Table II, transmission capacities vary within 2.9% of the `FLP` outcome, while a maximum of 4.1% Li-Ion storage capacity difference can be observed during the same year where distributed PV differed the most from the benchmark (i.e., 2017).

Finally, Table III summarizes the computational performance gains (relative to the `FLP`) achieved by leveraging the `SM`. More specifically, the reductions in i) the CEP problem size (number of variables, constraints and non-zeros), ii) the peak memory requirements (PMR) and iii) the solver runtime (or SRT, taking into account the solver runtime of both the `SITE` and `RLP` stages of the `SM`) are reported. In this table, it can be observed that the proposed `SM` leads to an average CEP problem size reduction of 33% which, in turn, enables an average PMR reduction of 40% and runtime savings between 31% and 46% across the studied instances.

V. CONCLUSION

This paper proposes a method to reduce the spatial dimension of CEP frameworks while preserving an accurate representation of renewable energy sources. This is achieved via a two-stage heuristic. First, a screening stage is used to identify the most relevant sites for RES deployment among a pool of candidate locations and discard the rest. Then, the subset of RES sites identified in the first stage is used in a CEP problem to determine the optimal power system configuration. The proposed method is tested on a realistic EU case study and its performance is assessed against a CEP set-up in which the entire set of candidate RES sites is available. The method shows great promise and manages to consistently identify more than 90% of the optimal sites while reducing peak memory consumption and solver runtime by up to 41% and 46%, respectively.

TABLE III: Computational performance assessment of the `SM`. Numerical values represent reductions associated with the `SM` expressed in relative terms (%) with respect to the `FLP`.

Year	Variables	Constraints	Non-Zeros	PMR	SRT
2016	34.54	34.67	33.48	41.37	36.56
2017	33.31	33.39	32.22	40.11	30.90
2018	33.72	33.82	32.67	39.28	46.57

Capacity differences between the solutions provided by the proposed method and the benchmark observed for some weather years suggest that further work on the selection of parameters used in the first-stage siting routine would be useful. Moreover, re-casting the proposed heuristic into a more structured form, e.g., where the siting and sizing of RES assets are used as stages in a Benders-like decomposition framework, is also envisaged as a promising development avenue.

REFERENCES

[1] S. Pfenninger, A. Hawkes, and J. Keirstead, "Energy systems modeling for twenty-first century energy challenges," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 74 – 86, 2014.

[2] A. MacDonald, C. Clack, A. Alexander, A. Dunbar, J. Wilczak, and Y. Xie, "Future cost-competitive electricity systems and their impact on US CO$_2$ emissions," *Nature Climate Change*, vol. 6, 2016.

[3] G. Wu, R. Deshmukh, K. Ndhlukulac, T. Radojicic, J. Reilly-Moman, A. Phadke, D. Kammen, and D. Callaway, "Strategic siting and regional grid interconnections key to low-carbon futures in African countries," *Proceedings of the National Academy of Sciences*, vol. 114, 2017.

[4] M. Hoffmann, L. Kotzur, D. Stolten, and M. Robinius, "A review on time series aggregation methods for energy system models," *Energies*, vol. 13, no. 3, 2020.

[5] NREL, "Regional Energy Deployment System (ReEDS) model documentation: version 2018," tech. rep., National Renewable Energy Laboratory, 2019. https://www.nrel.gov/docs/fy19osti/72023.pdf.

[6] J. Hörsch and T. Brown, "The role of spatial scale in joint optimisations of generation and transmission for European highly renewable scenarios," *2017 14th Intl. Conf. on the European Energy Market (EEM)*, 2017.

[7] B. A. Frew and M. Z. Jacobson, "Temporal and spatial tradeoffs in power system modeling with assumptions about storage: An application of the POWER model," *Energy*, vol. 117, pp. 198 – 213, 2016.

[8] European Centre for Medium-Range Weather Forecasts - ECMWF, "Copernicus knowledge base - ERA5 data documentation," 2018. https://confluence.ecmwf.int//display/CKB/.

[9] International Electrotechnical Commission, "IEC 61400-1:2019: Wind energy generation systems - Part 1: Design requirements." https://webstore.iec.ch/publication/26423, 2019.

[10] D. S. Ryberg, M. Robinius, and D. Stolten, "Methodological framework for determining the land eligibility of renewable energy sources," *arXiv preprint arXiv:1712.07840*, 2017.

[11] K. Freeman, C. Frost, G. Hundleby, A. Roberts, B. Valpy, H. Holttinen, L. Ramirez, and I. Pineda, "Our energy, our future - How offshore wind will help Europe to go carbon-neutral," tech. reps., WindEurope, 2019.

[12] T. Tröndle, S. Pfenninger, and J. Lilliestam, "Home-made or imported: On the possibility for renewable electricity autarky on all scales in Europe," *Energy Strategy Reviews*, vol. 26, p. 100388, 2019.

[13] Open Power System Data, "Data Package Time series." https://data.open-power-system-data.org/time_series/2019-06-05, 2019.

[14] ENTSO-E, "Maps & Data: TYNDP2018." https://tyndp.entsoe.eu/maps-data, 2018.

[15] European Commission - Joint Research Centre, "JRC Hydro-power plants database." https://github.com/energy-modelling-toolkit/hydro-power-database, 2020.

[16] T. Brown, J. Hörsch, and D. Schlachtberger, "PyPSA: Python for Power System Analysis," *Journal of Open Research Software*, vol. 6, 2018.

[17] A. Dubois, D. Radu, and M. Berger, "Model reduction in capacity expansion planning problems via renewable generation site selection - dataset," 2020. https://dox.uliege.be/index.php/s/errFWRnFbfroi17.

[18] A. Dubois, D. Radu, and M. Berger, "replan - a framework for bulk energy systems planning and analysis," 2020. https://github.com/montefesp/replan/releases/tag/v0.0.3.

[19] A. Malvaldi, S. Weiss, D. Infield, J. Browell, P. Leahy, and A. M. Foley, "A spatial and temporal correlation analysis of aggregate wind power in an ideally interconnected Europe," *Wind Energy*, vol. 20, no. 8, pp. 1315–1329, 2017.

[20] F. Neumann and T. Brown, "The near-optimal feasible space of a renewable power system model," *Electric Power Systems Research*, vol. 190, p. 106690, 2021.

NOMENCLATURE

A. Abbreviations

AC, DC	AC, DC transmission links
PV_u, PV_d	utility-scale PV, distributed PV
TSCE	total system cost error
W_{on}, W_{off}	onshore wind, offshore wind

B. Indices & Sets

$g, \mathcal{G}$ — conventional gen. tech. index and associated set

$l, \mathcal{L}$ — line, set of transmission corridors, $\mathcal{L} \subseteq \mathcal{N}_B \times \mathcal{N}_B$

$\mathcal{L}_n^+, \mathcal{L}_n^-$ — set of in-bound links into node n, with $\mathcal{L}_n^+ = \{l \in \mathcal{L} | l = (u, n), u \in \mathcal{N}_n^+\}$, where $\mathcal{N}_n^+ = \{u \in \mathcal{N}_B | (u, n) \in \mathcal{L}\}$ and set of out-bound links from node n, with $\mathcal{L}_n^- = \{l \in \mathcal{L} | l = (n, v), v \in \mathcal{N}_n^-\}$, where $\mathcal{N}_n^- = \{v \in \mathcal{N}_B | (n, v) \in \mathcal{L}\}$

$n, \mathcal{N}_B$ — bus, set of buses

$m, \mathcal{N}_R$ — candidate RES site and the associated set

$\mathcal{N}_R^n, \mathcal{N}_R^r$ — subset of sites assigned to bus $n \in \mathcal{N}_B$, subset of sites with RES tech. $r \in \mathcal{R}$

$\mathcal{N}_{SITE}^n$ — set of RES sites connected to bus $n \in \mathcal{N}_B$ retained in the siting stage

$\mathcal{N}_{SITE}^r$ — set of RES sites with tech. $r \in \mathcal{R}$ retained in the siting stage

$r, \mathcal{R}$ — tech. index and set of RES technologies

$s, \mathcal{S}$ — storage tech. and the associated set

$t, \mathcal{T}$ — time index, set of time periods

$\tau, \mathcal{T}_\tau$ — time slice index and time slice

C. Parameters

$\eta_s^{SD/D/C}$ — efficiencies of storage tech. s

ϕ_s — power-to-energy ratio of storage tech. s

ξ_τ^n — RES supply target at node n, in period τ

κ_l^0 — initial capacity for transmission line l

$\bar{\kappa}_l$ — maximum allowable installed capacity of line l

κ_{nx}^0 — initial capacity of tech. $x \in \{g, m, s\}$ at node n

$\bar{\kappa}_{nx}$ — maximum allowable installed capacity of tech. $x \in \{g, m, s\}$ at node n

λ_{nt} — electricity demand at node n and time t

π_{nmt} — per-unit availability of RES tech. at site m (connected to bus n) and time t

θ^e — economic penalty for demand curtailment

θ_f^x, θ_v^x — fixed (FOM) and variable (VOM) operation and maintenance cost of tech. $x \in \{g, m, s, l\}$

ω — weighting factor relating the length of time horizon $\mathcal{T}$ to the annualized costs of technologies

ζ^x — annualized investment cost of tech. $x \in \{g, m, s, l\}$

D. Variables

$K_l \in \mathbb{R}_+$ — installed capacity of transmission line l

$K_{nx} \in \mathbb{R}_+$ — installed cap. of tech. $x \in \{g, m, s\}$ at bus n

$p_{nst}^{C/D} \in \mathbb{R}_+$ — flows of storage tech. s at bus n and time t

$p_{lt} \in \mathbb{R}$ — power flow over line l at time t

$p_{nxt} \in \mathbb{R}_+$ — feed-in of gen. tech./site $x \in \{g, m\}$ at bus n and time t

$p_{nt}^e \in \mathbb{R}_+$ — unserved demand at bus n and time t