Hierarchical Clustering to Find Representative Operating Periods for Capacity-Expansion Modeling

Yixian Liu, Ramteen Sioshansi, *Senior Member, IEEE,* and Antonio J. Conejo, *Fellow, IEEE*

***Abstract*—Power system capacity-expansion models are typically intractable if every operating period is represented. This issue is normally overcome by using a subset of representative operating periods. For instance, representative operating hours can be selected by discretizing the load-duration curve, which captures the effect of load levels on system-operation costs. This approach is inappropriate if system-operating costs depend on parameters other than load (*e.g.*, renewable-resource availability) or if there are important intertemporal operating constraints (*e.g.*, generator-ramping limits).**

This paper proposes the use of representative operating days, which are selected using clustering, to surmount these issues. We propose two hierarchical clustering techniques, which are designed to capture the important statistical features of the parameters (*e.g.*, load and renewable-resource availability), in selecting representative days. This includes temporal autocorrelations and correlations between different locations. A case study, which is based on the Texan power system, is used to demonstrate the techniques. We show that our proposed clustering techniques result in investment decisions that closely match those made using the full unclustered data set.

***Index Terms*—Power system planning, representative days, hierarchical clustering, k-means clustering, dynamic time warping**

NOMENCLATURE

A. Clustering Variables and Functions

$\lvert\cdot\rvert$	number of points in a cluster or set.
$\mathcal{C}$	number of clusters.
C_i^k	cluster i in iteration k of clustering algorithm.
$\bar{C}$	centroid of the cluster, C.
$d(\mathbf{x},\mathbf{x}')$	distance between the points, $\mathbf{x}$ and $\mathbf{x}'$.
$d_{\mathbf{x},C}^{\max}$	maximal distance between the point, $\mathbf{x}$, and cluster, C.
$\mathcal{L}(C,C')$	minmax linkage between the clusters, C and C'.
n	dimension of points.
$r(C)$	minmax radius of the cluster, C.
S	a time series.
$U_{i,j}$	accumulated distance between element i of the time series, S, and element j of the time series, S'.
$\mathbf{x}$	a point.
X	a set of points.
$\Delta(S_i,S'_j)$	distance between element i of the time series, S, and element j of the time series, S'.

This work was supported by NSF grants 1029337 and 1548015.

Y. Liu and R. Sioshansi are with the Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH 43210, USA (e-mail: liu.2441@osu.edu, sioshansi.1@osu.edu).

A. J. Conejo is with the Department of Integrated Systems Engineering and the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA (e-mail: conejonavarro.1@osu.edu).

B. Operating-Period Data

ω_y	day-y operating-condition data.
ω_y^D	day-y demand data.
ω_y^S	day-y solar-insolation data.
ω_y^W	day-y wind-speed data.
$\omega_{e,y,h}^D$	hour-h load in region e on day y [MW].

C. Capacity-Expansion Model Sets

$\mathcal{G}$	set of generation technologies.
$\mathcal{H}$	set of hours in each day.
$\mathcal{E}$	set of regions.
$\mathcal{Y}$	set of representative operating days.

D. Capacity-Expansion Model Parameters

$\bar{B}_{g,e}$	maximum capacity of generation technology g that can be built in region e [MW].
$K_{g,e}^G$	operating cost of generation technology g in region e [$/MWh].
$K_{e,e'}^L$	investment cost of transmission between regions e and e' [$/MW].
K_e^S	investment cost of storage in region e [$/MW].
K^U	cost of unserved load [$/MWh].
$K_{g,e}^V$	investment cost of generation technology g in region e [$/MW].
$L_{e,y,h}$	region e's hour-h load on day y [MW].
δ_g	ramping factor of generation technology g [p.u.].
η	energy capacity of storage [hours of storage].
ζ	roundtrip efficiency of storage [p.u.].
$\phi_{g,e,y,h}$	hour-h capacity factor of generation technology g in region e on day y [p.u.].
Υ_y	weight on day y [days].

E. Capacity-Expansion Model Variables

$q_{e,y,h}^C$	hour-h power charged into storage in region e on day y [MW].
$q_{e,y,h}^D$	hour-h power discharged from storage in region e on day y [MW].
$q_{g,e,y,h}^G$	hour-h production from generation technology g in region e on day y [MW].
$q_{e,e',y,h}^L$	hour-h net power flow on link between regions e and e' on day y [MW].
$q_{e,y,h}^S$	hour-h ending state of charge of storage in region e on day y [MW].
$q_{e,y,h}^U$	hour-h unserved load in region e on day y [MW].
$z_{g,e}^G$	capacity of generation technology g built in region e [MW].

$z^L_{e,e'}$ transmission capacity built between regions e and e' [MW].
z^S_e storage capacity built in region e [MW].

F. Clustered and Unclustered Data

$z^C_{\tau,\gamma}$ investment in technology τ using clustered data when wind-investment costs are $\gamma\%$ below baseline.
$z^U_{\tau,\gamma}$ investment in technology τ using unclustered data when wind-investment costs are $\gamma\%$ below baseline.
Γ set of cases with different wind-investment costs.
$\rho^{C,W}_{e,i}$ ith element of wind-duration curve in region e in clustered data [p.u.].
$\rho^{U,W}_{e,i}$ ith element of wind-duration curve in region e in unclustered data [p.u.].

I. Introduction

HOUR-to-hour solar and wind availabilities and demand are important uncertain and variable factors in power system operations and capacity expansion. These factors exhibit multiple important correlations. First, wind, solar, and demand may generally be correlated. In many systems demand is low during the night when wind speeds are high and solar insolation is zero. Second, each variable is autocorrelated. Considering autocorrelation helps with modeling intertemporal operating constraints, such as generator-ramping limits. Third, there are spatial correlations when multiple locations are considered. For example, the wind speed may be low at one location but simultaneously high elsewhere. Not capturing spatial correlation may result in misrepresenting the impacts of renewable generation.

Capacity-expansion models capture investment and operating decisions. Investment decisions are typically made at coarse time intervals (*e.g.*, yearly or decennially). Conversely, operating decisions are made at finer time scales (*e.g.*, hourly or sub-hourly). Therefore, the operating period must be represented many more times than the investment period, resulting in a computationally challenging problem. Several works surmount this issue by using a reduced set of operating periods.

The pioneering work of Caramanis *et al.* [1] accounts for the impact of non-dispatchable resources on the system load profile. They employ a stochastic approach to modify the yearly load prior to selecting a number operating conditions from the modified load. Short *et al.* [2] develop a deterministic capacity-expansion model that captures seasonal and diurnal variability in demand and resource profiles using 17 time-slices. Each season is represented by one day, each of which is represented by four time-slices. There is also one summer super-peak time slice. This representation is mainly based on demand patterns and its effectiveness in representing the complete hourly data is not studied. Pina *et al.* [3] divide a year into four seasons, each of which is represented by three days that are modeled at hourly resolution. Baringo and Conejo [4] propose two methods to find representative hours. The first uses load- and wind-duration curves. This technique cannot model spatial correlations, however, which is overcome by their second method, which employs k-means clustering. However, their use of representative hours breaks the chronological sequence of the operating stages and cannot represent intertemporal operating constraints. Wogrin *et al.* [5] take a markedly different approach, wherein they use system states as opposed to load levels to characterize operating conditions in a capacity-expansion model. They claim that the system states that they define embody more operational information than loads alone do. Poncelet *et al.* [6] develop a metric to select representative days in expansion planning exercises in systems with storage and illustrate the use of their proposed metric. Alvarez *et al.* [7] provide a technique to select representative operating conditions for transmission-expansion planning. Their method focuses on critical network conditions, as opposed to modeling 'more common' nominal conditions. Ploussard *et al.* [8] use a snapshot selection technique to capture a suitable number of operating conditions for transmission-expansion planning.

Using representative operating periods to reduce the complexity of capacity-expansion models is desirable. However, many techniques in the literature do not provide a sound basis on which to select operating periods, as our results illustrate. Instead, most select them in an *ad hoc* manner, such as selecting a fixed number of periods from each season. Moreover, the techniques historically employed have difficulty in capturing all of the relevant correlations in the data. Finally, most of the techniques used break the chronological sequence of the data, complicating the modeling of intertemporal operating constraints.

In this paper, we propose using representative days (as opposed to hours or time-slices) to represent operating decisions in capacity-expansion models. Doing so allows modeling intertemporal operating constraints. We study two clustering techniques, which capture the relevant correlations, to select representative days. The first employs hierarchical clustering while the second consists of k-means clustering followed by hierarchical clustering within each k-means cluster. We demonstrate the effectiveness of our proposed techniques in two ways using a case study based on the Texas power system. First, we compare the properties of the clustered and unclustered data sets in terms of capturing the range of different renewable-resource and load conditions. Second, we show that the optimized investments made using the representative days very closely match investments made using a full year's hourly data in the operating stage.

The clustering techniques that we employ are not, in and of themselves, novel. Rather, the novelty of our work is in developing a mixture of clustering methods, linkage criteria, and distance metrics (*cf.* Section II-B for definitions of these latter two terms) that provide good performance in efficiently determining a set of representative operating periods for capacity-expansion modeling. Our work focuses on selecting representative days. This focus is motivated by the desire to represent generator-ramping constraints and intraday energy storage in capacity-expansion modeling. These features cannot be represented properly without having a temporal sequence of operating periods, as given by representative operating days. Liu *et al.* [9] demonstrate the importance of representing generator-ramping constraints in ensuring that a system's

generation mix has sufficient flexibility to deal with variability in load and renewable production. Nevertheless, representative operating weeks or longer-duration operating periods may be more appropriate if, for instance, interday or seasonal energy storage is an important feature of a capacity-expansion model. The methods that we propose could be employed to select such operating periods.

The remainder of this paper is organized as follows. Section II discusses commonly used clustering methods and details our two proposed methods. Section III gives the formulation of the capacity-expansion model that is used in testing our clustering methods. Sections IV and V summarize the case study conducted and its results, respectively. Section VI concludes.

II. Clustering Techniques

The basic objective in cluster analysis is to discover natural groupings of items based on either their similarities or dissimilarities. k-means and hierarchical clustering are both very popular clustering methods with strengths and weaknesses. We begin in this section by first introducing some basic concepts of k-means and hierarchical clustering in Sections II-A and II-B, respectively. We assume throughout this discussion that the different 'items' being clustered are represented as vectors, which can specify the features of the items on multiple dimensions. We then detail our two proposed clustering techniques in Section II-C.

A. k-Means Clustering

k-means clustering assigns each item within a data set to a cluster that has its centroid closer to the item than the centroid of any other cluster [10]. The primary benefit of k-means clustering is that it performs relatively quickly compared to other clustering methods. On the other hand, k-means clustering requires some restrictive assumptions on the data being clustered. Importantly, to be effective k-means clustering requires that the data be in similarly sized hyperspherical clusters [11]. Algorithm 1 summarizes the most commonly used variant of the k-means clustering algorithm [12], which takes two inputs—a set of points (or vectors) to be clustered and the number of clusters to assign the points to. Step 2 initializes the algorithm by setting the iteration counter to 0 and then assigning each point to one of $\mathcal{C}$ starting clusters. The final set of clusters obtained is generally dependent on this initial assignment of points to clusters.

Steps 3–13 are the main iterative loop. Step 5 recomputes the centroid of each cluster and Step 6 initializes each cluster in the next iteration to be empty. Next, in Step 9, a cluster with the nearest centroid to each point is determined and each point is assigned to a cluster with the nearest centroid in Step 10. Although any distance metric can be used, Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1}^{n} (\mathbf{x}_j - \mathbf{x}'_j)^2}, \tag{1}$$

is quite common. This iterative process repeats until no reassignments are done between two successive iterations (*cf.* Step 13).

Algorithm 1 k-Means Clustering

1: **inputs**: $X, \mathcal{C}$
2: **initialize**: $k \leftarrow 0$; assign each point to starting clusters, $C_1^0, \ldots, C_{\mathcal{C}}^0$
3: **repeat**
4: **for** $i \leftarrow 1, \ldots, \mathcal{C}$ **do**
5: $\bar{C}_i \leftarrow \left(\sum_{\mathbf{x} \in C_i^k} \mathbf{x}\right) / |C_i^k|$
6: $C_i^{k+1} \leftarrow \emptyset$
7: **end for**
8: **for** $\mathbf{x} \in X$ **do**
9: $i \leftarrow \arg \min_{i' \in \{1,\ldots,\mathcal{C}\}} d(\mathbf{x}, \bar{C}_{i'})$
10: $C_i^{k+1} \leftarrow C_i^{k+1} \cup \mathbf{x}$
11: **end for**
12: $k \leftarrow k + 1$
13: **until** $C_i^k = C_i^{k-1}, \forall i = 1, \ldots, \mathcal{C}$

B. Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters using either a 'bottom-up' or 'top-down' approach [13]. The former, agglomerative, approach begins with each single point as its own cluster. The algorithm proceeds by successively merging clusters until attaining the desired number. The latter, divisive, approach begins with all points in a single cluster, which are successively divided. The result of both approaches can be displayed as a dendrogram, which illustrates the successive mergers or divisions.

Unlike k-means clustering, hierarchical clustering has the benefit that it can be applied to data with clusters that are not hyperspherical [11]. Moreover, the final set of clusters obtained are not dependent on the initial allocation of points to clusters. These benefits come at computational and data-storage costs, because hierarchical clustering requires computing and storing a matrix of distances between all sets of clusters. As a result, k-means clustering can typically be applied to much larger data sets compared to hierarchical clustering [12].

Hierarchical clustering requires a measure of dissimilarity between sets of observations, which are stored as a matrix of distances, to determine mergers or divisions of clusters. Thus, a linkage criterion and a distance metric must be specified. We now discuss the linkage criterion and distance metric that are used in our proposed clustering methods.

1) Linkage Criterion: The linkage criterion measures the distance between two clusters. Commonly used linkage criteria include single, complete, and average linkage. Single linkage determines the distance between two clusters based on the distance between the two elements (one in each cluster) that are closest to one another. Complete linkage is based, conversely, on the maximum distance between two elements of the clusters. Average linkage is based on the average distance between pairs of elements of two clusters.

Our proposed clustering techniques employ minmax linkage [14], [15]. This linkage criterion determines a point, which is referred to as the cluster prototype, which can be thought of as a point within the cluster that is most representative of

it. Having cluster prototypes is beneficial, because they allow each cluster to be represented by its prototype in modeling system operations in the capacity-expansion model.

To compute the minmax linkage between two clusters, we first define the maximal distance between any point, $\mathbf{x}$, and the cluster, C, as:

$$d^{\max}_{\mathbf{x},C} = \max_{\mathbf{x}' \in C} d(\mathbf{x}, \mathbf{x}'). \tag{2}$$

In words, $d^{\max}_{\mathbf{x},C}$ is defined as the distance between $\mathbf{x}$ and the point in C that is furthest from $\mathbf{x}$. We then define the minmax radius of the cluster, C, as:

$$r(C) = \min_{\mathbf{x} \in C} d^{\max}_{\mathbf{x},C}. \tag{3}$$

The point, $\mathbf{x}_C$, that minimizes (3) is defined at the prototype of the cluster, C. The prototype has the property that it has the minimal maximal distance to C. From these definitions, we also know that a closed ball of radius, $r(C)$, centered at the prototype covers all of the points in C. We finally define the minmax linkage between the clusters C and C' as:

$$\mathcal{L}(C, C') = r(C \cup C'). \tag{4}$$

Minmax linkage defines the distance between two clusters by the minmax radius of the union of the clusters. A larger minmax radius means that a larger ball is needed to cover all of the points in the union of the clusters.

2) Distance Metric: Both k-means and hierarchical clustering require a measure of similarity (*i.e.*, a distance metric). Choosing a distance metric when clustering time-series data adds a complication. This is because standard distance metrics align the jth element in one time series with the jth element of another. This is seen, for instance, in (1), which defines Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$. Thus, standard distance metrics assume that the time-series data are aligned on the time axis.

In practice, time series may not be perfectly aligned or may have other shape properties, which could result in a poor similarity measure. For instance, there is a one-hour offset in load and other data between days during daylight-savings and standard times. As another example, utilities and other entities may use different practices in recording load and other data at the beginning, middle, or end of the corresponding timestamp [16]. As a final example, two days may have similar patterns in terms of the magnitude of the peaks and troughs in load or other data. However, the time at which the peaks and troughs occur may differ. In these and other cases, a distance metric that assumes that the time-series data are aligned on the time scale, such as Euclidean distance, may give poor fits between days that are actually similar.

Dynamic time warping (DTW) is a distance metric that addresses this issue in measuring the similarity of time series [17]. DTW measures the similarity between two time series by computing the optimal (least cumulative distance) alignment between their elements. In doing so, DTW allows time series with similar shapes to match even if they are out of phase in the time axis. Thus, DTW produces a more suitable measure of similarity for time series.

To illustrate how DTW measures the distance between two generic time series, let $S = \{S_1, \ldots, S_I\}$ and $S' = \{S'_1, \ldots, S'_J\}$ be two time series, which can have different lengths. We also let $\Delta(S_i, S'_J)$ denote a measure of distance between element i of S and element j of S'. One can use $\Delta(S_i, S'_j) = |S_i - S'_j|$, or a different distance measure.

Algorithm 2 summarizes the steps of a commonly used variant of the DTW algorithm [17], which takes two time series as inputs. The ultimate output of the algorithm is a matrix, U, with cumulative distances between different elements of the two time series, taking into account the amount that the time axes of the two series are warped. Steps 2–7 initialize the U matrix by setting the cumulative distance equal to $+\infty$ if either time series index equals zero. The actual cumulative distances are calculated in Step 10. The first term, $\Delta(S_i, S'_j)$, measures the distance between the ith and jth elements of the two time series. The second term takes into account the added cumulative distance (up to elements i and j of the two time series), depending on which (if either) of the two time axes are warped [17]. The 'warping' of the time axes in represented by the $U_{i-1,j}$ and $U_{i,j-1}$, which allow a 'mismatch' in the time axes of the two time series. The distance between the two time series is given by the final value of $U_{I,J}$, which measures total distance between the two time series (taking into account any warping of the time axes), after the algorithm terminates.

Algorithm 2 Dynamic Time Warping Algorithm

1: **inputs**: S, S'
2: **for** $i \leftarrow 1, \ldots, I$ **do**
3: $\quad U_{i,0} \leftarrow +\infty$
4: **end for**
5: **for** $j \leftarrow 1, \ldots, J$ **do**
6: $\quad U_{0,j} \leftarrow +\infty$
7: **end for**
8: **for** $i \leftarrow 1, \ldots, I$ **do**
9: $\quad$ **for** $j \leftarrow 1, \ldots, J$ **do**
10: $\quad\quad U_{i,j} \leftarrow \Delta(S_i, S'_j) + \min\{U_{i-1,j}, U_{i,j-1}, U_{i-1,j-1}\}$
11: $\quad$ **end for**
12: **end for**

C. Proposed Clustering Methods

Our proposed clustering methods are rooted in two requirements. First, the method should generate representative days that respect all of the important correlations in the data. Second, the clustering method should be able to group large data sets quickly. To meet the first requirement, we cluster electricity demand and renewable-availability data for each region or node that is represented in the capacity-expansion model.

To illustrate our proposed clustering techniques, we first define:

$$\omega^D_y = \left(\omega^D_{1,y,1}, \ldots, \omega^D_{1,y,24}, \omega^D_{2,y,1}, \ldots, \omega^D_{|\mathcal{E}|,y,24}\right), \tag{5}$$

$\forall y = 1, \ldots, 365$ as a vector of day-y demand data for all of the hours and regions. We similarly define:

$$\omega^S_y = \left(\omega^S_{1,y,1}, \ldots, \omega^S_{1,y,24}, \omega^S_{2,y,1}, \ldots, \omega^S_{|\mathcal{E}|,y,24}\right), \tag{6}$$

$\forall y = 1, \ldots, 365$ and:

$$\omega_y^W = \left(\omega_{1,y,1}^W, \ldots, \omega_{1,y,24}^W, \omega_{2,y,1}^W, \ldots, \omega_{|\mathcal{E}|,y,24}^W\right), \quad (7)$$

$\forall y = 1, \ldots, 365$ as vectors of day-y solar- and wind-availability data, respectively. The load, solar, and wind observations in these three vectors are ordered first by hour of day and then by location modeled. This is to ensure that the temporal sequence of observations is not lost. We also define:

$$\omega_y = \left(\omega_y^D, \omega_y^S, \omega_y^W\right) \quad (8)$$

$\forall y = 1, \ldots, 365$ as a vector of day-y operating-condition data.

By definition, ω_y contains all of the pertinent data on which to cluster (which in our case study are wind and solar conditions and load) for each region represented in the capacity-expansion model. As such, clustering on the collection of points, $\omega_1, \ldots, \omega_{365}$, yields a set of days that capture a wide variety of operating (*i.e*., wind, solar, and load) conditions at the different regions modeled. These operating days fully respect the intraday serial correlation in operating conditions, because the time sequence of operating data are maintained in the final set of operating conditions. Interregional correlations are also captured, because each representative day contains contemperaneous operating-condition data for all of the regions that are modeled. Although interday correlations are not captured by the representative days, interday and seasonal variability in operating conditions are. This is because clustering on $\omega_1, \ldots, \omega_{365}$ will yield a set of representative days that 'differ' from one another in terms of their operating conditions.

Our two proposed clustering methods differ in terms of achieving the second design requirement. The first proposed clustering method applies agglomerative hierarchical clustering with a minmax linkage criterion and DTW distance metric to the full set of operating-condition data, $\omega_1, \ldots, \omega_{365}$. This method, which we hereafter refer to as HC, has all of the benefits of hierachical clustering in not requiring similarly sized hyperspherical clusters. Moreover, the use DTW as a distance metric helps to provide more robust clusters. For instance, two days may have similar load and renewable-availability conditions, but the exact patterns may be out of time phase (*e.g*., due to daylight-savings time). DTW helps with controlling for such time-phase issues. On the other hand, the HC method may scale poorly (*i.e*., as more regions or underlying operating-condition data are considered).

The second method first applies k-means clustering to the full set of operating-condition data, $\omega_1, \ldots, \omega_{365}$, to obtain a starting set of clusters, $C_1, \ldots, C_{\mathcal{C}}$. We then apply the same agglomerative hierarchical clustering technique that is used in HC to the days within each k-means cluster, $C_1, \ldots, C_{\mathcal{C}}$. The benefit of this second method, which we hereafter refer to as kMHC, is that it first uses the relatively fast k-means clustering algorithm to obtain an initial set of clusters. Slower hierarchical clustering is then applied within each k-means cluster to obtain the final set of clusters, retaining some of the benefits of the HC method.

In both methods, after the final set of clusters is obtained, the days in each cluster are represented in the capacity-expansion model by the cluster prototype. The weight placed on each cluster prototype is equal to the number of days within the corresponding cluster.

III. Capacity-Expansion Model

We use a capacity-expansion model to analyze how the representative days that are selected by the proposed clustering techniques affect investment decisions. The model that we use is a simplification of the multistage, multiscale stochastic capacity-expansion model that is proposed by Liu *et al*. [9]. Whereas the model of Liu *et al*. is stochastic and has multiple investment periods, the model that we employ is static, linear, and deterministic. This is to simplify the model structure and allow us to solve it with the full unclustered data for purposes of comparing model results to using clustered data. Nevertheless, the clustering methods that we propose could be used to select representative operating days within a stochastic capacity-expansion model [18].

The model that we use assumes that a single set of investment decisions is first made. These are then followed by hourly operating decisions over twenty years. To simplify the capacity-expansion model, all of the investments are assumed to be continuous and unit commitment decisions are not considered at the operating stage. The transmission network is represented using a pipeline model. Moreover, we do not include planning-reserve or other types of reliability-related constraints.

The capacity-expansion model is formulated as:

$$\min \sum_{e\in\mathcal{E}} \left[\sum_{g\in\mathcal{G}} K_{g,e}^V z_{g,e}^G + \sum_{e'\in\mathcal{E}, e'\neq e} K_{e,e'}^L z_{e,e'}^L + K_e^S z_e^S + \sum_{y\in\mathcal{Y}, h\in\mathcal{H}} \Upsilon_y \cdot \left(\sum_{g\in\mathcal{G}} K_{g,e}^G q_{g,e,y,h}^G + K^U q_{e,y,h}^U \right) \right] \quad (9)$$

$$\text{s.t. } 0 \le z_{g,e}^G \le \bar{B}_{g,e}, \quad \forall g, e \quad (10)$$

$$0 \le z_{e,e'}^L, \quad \forall e, e' \neq e \quad (11)$$

$$0 \le z_e^S, \quad \forall e \quad (12)$$

$$\sum_{g\in\mathcal{G}} q_{g,e,y,h}^G + \sum_{e'\in\mathcal{E}, e'\neq e} \left(q_{e',e,y,h}^L - q_{e,e',y,h}^L\right) + q_{e,y,h}^D - q_{e,y,h}^C + q_{e,y,h}^U = L_{e,y,h}, \quad \forall e, y, h \quad (13)$$

$$0 \le q_{g,e,y,h}^G \le \phi_{g,e,y,h} z_{g,e}^G, \quad \forall g, e, y, h \quad (14)$$

$$-\delta_g z_{g,e}^G \le q_{g,e,y,h}^G - q_{g,e,y,h-1}^G \le \delta_g z_{g,e}^G, \quad \forall g, e, y, h \quad (15)$$

$$-z_{e,e'}^L \le q_{e,e',y,h}^L \le z_{e,e'}^L, \quad \forall e, e' \neq e, y, h \quad (16)$$

$$q_{e,y,h}^S = q_{e,y,h-1}^S - q_{e,y,h}^D + \zeta q_{e,y,h}^C, \quad \forall e, y, h \ge 2 \quad (17)$$

$$q_{e,y,0}^S, q_{e,y,|\mathcal{H}|}^S = \frac{1}{2}\eta z_e^S, \quad \forall e, y \quad (18)$$

$$0 \le q_{e,y,h}^S \le \eta z_e^S, \quad \forall e, y, h \quad (19)$$

$$0 \le q_{e,y,h}^C, q_{e,y,h}^D \le z_e^S, \quad \forall e, y, h \quad (20)$$

$$0 \le q_{e,y,h}^U \le L_{e,y,h}, \quad \forall e, y, h. \quad (21)$$

Objective function (9) minimizes total cost, which consists of generation-, transmission-, and storage-investment costs, generator-operating cost, and the cost of unserved load. Investment costs can include the costs of constructing, maintaining, and eventually retiring assets. Generator-operating costs can include the costs of fuel and maintaining plants.

The model has two types of constraints. Constraints (10)–(12) pertain to investments whereas constraints (13)–(21) concern operations. Constraints (10) imposes limits on generation technology investments, for instance due to land restrictions, resource availability, or policy restrictions. Constraints (11) and (12) impose non-negativity on transmission and storage investments.

Constraints (13) impose load-balance in each hour. Constraints (14) and (15) impose capacity and ramping limits on generators. The capacity limits are defined based on total installed capacity available multiplied by a capacity factor. The capacity factor captures hour-to-hour variability in wind, solar, and other renewable availability, which is embedded in the representative days selected by clustering. The ramping limit is assumed to be a multiple of the installed capacity, with higher values of δ_g denoting a more flexible generating technology. Constraints (16) impose transmission-capacity limits.

Constraints (17)–(20) concern storage operations. Constraints (17) define the ending state of charge of storage in each hour. Constraints (18) force each storage device to begin and end each day with a 50% state of charge. This is a heuristic approach to attaching carryover value to stored energy from one day to the next [19]. Constraints (19) and (20) impose energy and power limits on storage. The energy capacity of storage is measured by the number of hours of full power output [20]. We assume that the storage technology modeled has no effective ramping limit. This is because many storage technologies in use today have no effective ramping limit.

Constraints (21) limit the amount of unserved energy in each operating period to be no greater than demand.

IV. Case Study Data

Our case study is based on the state of Texas, which is represented as consisting of three regions in the capacity-expansion model. Hourly solar-insolation, wind-speed, and temperature data are generated for each region using a vector autoregression model, which is calibrated to 16 years of hourly weather observations [21]. Solar-insolation and wind-speed data are input to models that estimate photovoltaic and wind-turbine outputs. The temperature data are used to simulate hourly residential, commercial, and industrial electricity demand data [22], [23]. Historical weather data can also be used in place of a regression model.

Total simulated demand for the state, assuming 2010 population levels, peaks in July at about 69 GW. The historical peak between 2006 and 2015 ranges between 62 GW and 70 GW and occurs in July or August, showing that our models capture load patterns well. The East region has higher demand than the other two regions, which is in keeping with actual system loads. Wind capacity factors are highest in the West region, which is also consistent with actual weather patterns. Finally, solar insolation peaks in the summer. However, solar capacity factors are not significantly higher in the summer compared to the winter, because cell temperatures are higher in the summer, reducing cell efficiency. Table I summarizes the mean values of the three operating-condition parameters, which we use in our cluster analysis, for the three regions.

TABLE I
Average Simulated Demand and Wind and Solar Capacity Factors in Three Regions in 2010

Region	Demand [MWh]	Wind [p.u.]	Solar [p.u.]
East	20408	0.40	0.19
West	8102	0.50	0.23
South	12226	0.46	0.20

Load and renewable capacity factors are reported in different units. As such, they must be normalized so that the calculated distance that is used in the clustering methods places the same weight on the three data sets. We normalize each of the load and renewable capacity-factor data for each location to have mean zero and standard deviation 1. Although we normalize load data for purposes of clustering, we report load data throughout this paper in absolute terms (*cf.*, for instance, Table I). This is because while wind and solar available are more naturally reported in p.u., load is not.

Our capacity-expansion model assumes a 'greenfield' system, with no starting generation, storage, or transmission capacity. This is because a brownfield system would see relatively small incremental investments, which would make it difficult to draw conclusions as to whether the clustering techniques yield representative operating days that result in good investment decisions. We consider five generic generation technologies: wind, solar, coal, natural gas, and nuclear. The model can also build a generic storage technology, which has $\eta = 20$ hours of storage capacity and a roundtrip efficiency of $\zeta = 0.8$. Table II summarizes the baseline technology-related parameters used in the model, which are obtained from the United States Energy Information Administration's 2014 Annual Energy Outlook [24] and other sources [25], [26]. Although these data sources are a few years old, they represent credible data sources that have been used in numerous United States Department of Energy technical studies. Moreover, we do not expect that the performance of the different clustering methods would be unduly affected by using different costs. The cost of load curtailment is assumed to be \$5000/MWh.

TABLE II
Baseline Parameter Values of Capacity-Expansion Model

Technology	Investment Cost [$/kW]	Operating Cost [$/MWh]	Ramp Rate [p.u.]
Wind	3737–3864	0	n/a
Solar	3164–3345	0	n/a
Coal	3037–3164	25–26	0.29
Natural Gas	837–964	44–45	0.43
Nuclear	6533–6562	9–10	0.16
Storage	2333-2362	n/a	n/a
Transmission	503-806	n/a	n/a

V. Case Study Results

We compare the performance of our two proposed clustering techniques in terms of selecting representative operating days and the resulting investments made by the capacity-expansion model. Our numerical testing suggests that a minimum of 30 representative days is needed to capture the range of load, wind, and solar conditions in capacity-expansion modeling [18]. For brevity we only present results for cases using 30 representative days here. Using more days results in higher-fidelity capacity-expansion modeling, but at higher computational cost. We generate 30 representative days with the kMHC technique by first applying k-means clustering to obtain 10 initial clusters. We then apply HC within each of those 10 clusters to find three subclusters. We also compare our proposed clustering techniques to using k-means clustering *only* to find 30 clusters.

A. Representative Days Selected

Fig. 1 summarizes the 30 days that are selected and the weights that are placed on them by the HC and kMHC techniques. The days selected are the prototypes of the clusters and the associated weights are the number of days within the clusters. The figure shows that both techniques select days in a non-uniform manner. For instance, the HC method selects eight days in the month of July but only two days in the first three months of the year. Moreover, the first three months of the year are only given 49 days of weight, which is less than 14% of the total. These results suggest that capacity-expansion models that apply uniform weights to each season of the year [2], [3] are somewhat arbitrary in nature, as we note in Section I.

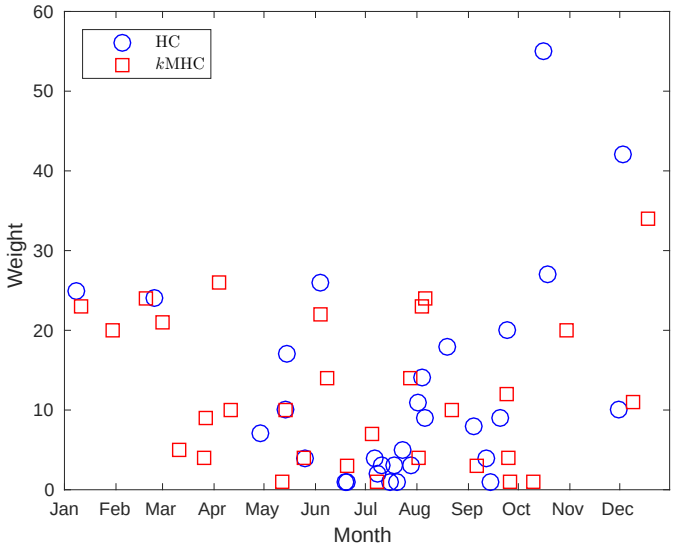

Fig. 1. Weights placed on 30 representative days by HC and kMHC methods.

Overall, the HC method selects days that are more non-uniform across the year relative to kMHC. This is in part because of the two-step clustering process underlying kMHC. Among the 30 days that they select, nine are common to both methods: days 134, 155, 171, 189, 209, 124, 126, 218, and 267. Moreover, HC and kMHC assign total weights of 96 and 94, respectively, to these nine days (which is about 26% of the total weight), showing further similarities between the two methods.

The nine commonly chosen days are mostly during the summer season, demonstrating the importance of the summer in capturing greater variability in different load and renewable-availability conditions. Table III summarizes the per-season standard deviation of load and wind and solar capacity factors in the unclustered data. The table reveals two interesting findings. First, it shows that there is considerably more variability in load as opposed to in renewable availability, when comparing across seasons. Secondly, the table shows that the summer season tends to see greater variability in operating conditions (compared to the other three seasons). Taken together, these two observations explain more weight being placed on the summer, as is shown in Fig. 1, and that load variability has an outsize impact on the weighting toward the summer.

TABLE III
PER-SEASON STANDARD DEVIATIONS OF LOAD AND WIND AND SOLAR CAPACITY FACTORS IN UNCLUSTERED DATA

	Winter	Spring	Summer	Fall
Load [GW]	5.46	8.02	10.50	7.90
Solar [p.u.]	0.74	0.74	0.76	0.76
Wind [p.u.]	0.74	0.74	0.72	0.73

Table IV shows the peak loads in the unclustered and clustered data. The table shows that both the HC and kMHC methods outperform k-means clustering in capturing the load peaks. While HC and kMHC come within 97% of the overall peak system load in the unclustered data, k-means clustering comes within 94% of this peak. Moreover, both the HC and kMHC methods outperform k-means clustering in capturing zonal peak loads. The HC method also captures the peak in two of the load zones.

TABLE IV
PEAK LOADS IN UNCLUSTERED AND CLUSTERED DATA [GW]

Zone	Unclustered Data	HC	kMHC	k-Means Clustering
East	35.24	35.11	35.11	33.64
South	21.09	21.09	19.60	19.43
West	13.85	13.85	12.84	12.54
Total	69.29	66.77	66.77	65.52

We cannot show which days are selected if only k-means clustering is used. This is because k-means clustering does not provide a representative day (*i.e.*, prototype) for each cluster. Instead, we represent each cluster found by the k-means algorithm by its centroid (*cf.* Step 5 of Algorithm 1).

Fig. 2 shows the wind-duration curve for the East region using the unclustered and clustered data. The figure shows mismatches, however the HC and kMHC methods outperform k-means clustering in representing the unclustered wind-duration curve. k-means clustering gives a worse match to the unclustered data because it 'over-averages' wind conditions. This is seen in the tails of the wind-duration curve. The representative days given by k-means clustering underestimate

wind production during high-wind days and overestimate wind production during low-wind days. This finding that k-means clustering over-averages renewable availability carries over to other resource/region combinations. We do not show other wind- and solar-duration curves for sake of brevity. This issue with k-means clustering can be overcome if more representative days are generated. However, this results in a larger and more computationally challenging capacity-expansion model.

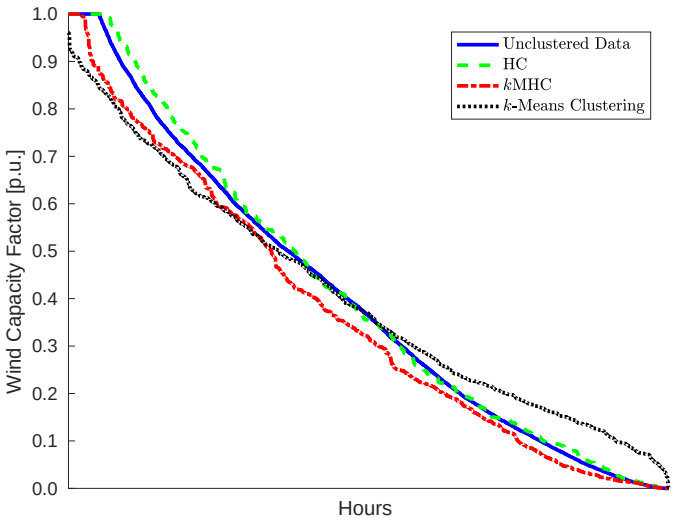

Fig. 2. East-region wind-duration curve using unclustered and clustered data.

We can quantify the match between the wind profiles given by the clustered and unclustered data using a normalized root mean square deviation (RMSD). We define the RMSD between the clustered and unclustered wind-duration curves as:

$$\frac{\sqrt{\frac{1}{8760}\sum_{i=1}^{8760}\left(\rho_{e,i}^{C,W}-\rho_{e,i}^{U,W}\right)^2}}{\frac{1}{8760}\sum_{i=1}^{8760}\rho_{e,i}^{U,W}}, \tag{22}$$

where the RMSD is normalized by the average wind capacity factor. The same metric can also be applied to the clustered and unclustered load- and solar-duration curves. In the case of the wind-duration curves for the East region, the HC and kMHC methods have RMSDs of 0.02 and 0.04, respectively, whereas k-means clustering has an RMSD of 0.16.

Table V summarizes the average (over the three locations modeled) RMSDs for the load-, solar-, and wind-duration curves given by the HC, kMHC, and k-means clustering techniques. HC and kMHC tend to outperform k-means clustering, which overly averages extreme conditions. This is especially true for renewable-resource data, because the three techniques perform the same in representing the load-duration curves. The RMSDs for the load-duration curves that are obtained from the three clustering techniques are nearly identical for two reasons. First, loads display less variability as compared to wind and solar availabilities. As such, the load-duration curves from the three clustering techniques are very similar. Secondly, load values are relatively large in magnitude. As such, what differences there are in the load-duration curves are negligible when normalized by the average load. HC slightly outperforms kMHC in terms of goodness-of-fit of wind-duration curves, whereas kMHC outperforms HC in terms of solar-duration curves. This suggests that the two methods are comparable in terms of selecting reasonable operating days.

TABLE V
AVERAGE (AMONG THE THREE REGIONS) NORMALIZED RMSD FOR LOAD, SOLAR, AND WIND-DURATION CURVES GIVEN BY HC, kMHC, AND k-MEANS CLUSTERING

	Demand	Solar	Wind
HC	1.02	0.11	0.06
kMHC	1.02	0.08	0.07
k-Means Clustering	1.02	0.14	0.13

The clustering methods are implemented in R on a system with a 3.10 GHz Intel Core i7-3770S processor and 8 GB of memory. The HC and kMHC methods take approximately 15 and two minutes of CPU time, respectively, to provide 30 clusters, while k-means clustering takes several seconds. This shows a further benefit of kMHC. The time difference between the HC and kMHC is expected to increase if the data sets being clustered grow in size. This could occur because more locations are modeled, more operating-condition data are used to select representative days, or the operating-condition data are recorded at subhourly intervals.

B. Investment Decisions

We test the effectiveness of the representative days in capacity-expansion modeling by examining cases in which wind-investment costs are reduced relative to the baseline values that are given in Table II. The rationale behind this analysis is that changes in the investment cost of wind will result in differences in the generation mix installed. For instance, lower wind-investment costs tends to increase wind investments. At the same time, greater wind investments may also call for changes in the mix of other generation resources, such as more flexible generation with greater ramping capabilities.

Figs. 3–5 summarize the investments in coal- and natural gas-fired generation and wind (aggregated across the three regions) by the capacity-expansion model with different wind-investment costs and representative operating days. There are no investments in nuclear or solar technologies in any of the cases that are examined. This is because of the high costs of nuclear and solar and the relatively low capacity factor of solar.

The figures show some trends in the investments, which are observed when using unclustered and clustered data sets. Reductions in wind-investment costs generally result in more wind generation being built in place of coal- and natural gas-fired capacity. This is because lower wind-investment costs means that wind is a lower-cost source of energy.

The one exception to this is for relatively modest reductions in wind-investment costs (*i.e.*, the case with a 30% reduction relative to baseline). In this case, wind and natural gas-fired capacities both increase, while coal-fired capacity decreases. The reason for this is that the greater wind capacity built requires more dispatchable and flexible natural gas-fired capacity (natural gas-fired generation has a relatively

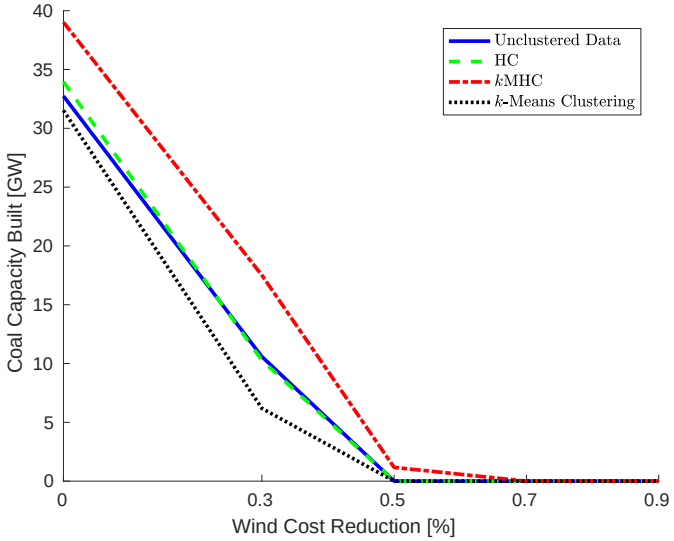

Fig. 3. Coal-fired capacity that is built as a function of the reduction in wind-investment cost relative to baseline using unclustered and clustered data.

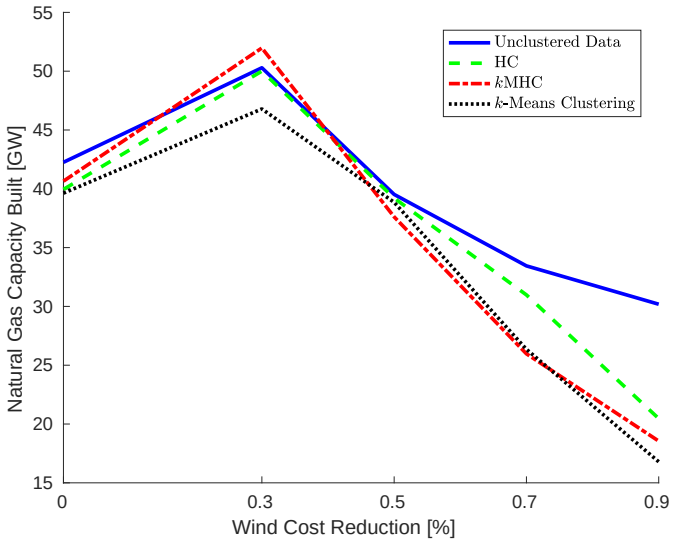

Fig. 4. Natural gas-fired capacity that is built as a function of the reduction in wind-investment cost relative to baseline using unclustered and clustered data.

high ramping capability). This observation is consistent with other analyses of the potential impacts of high penetrations of renewable energy [27]. Further reductions in wind-investment costs beyond the 30% case allow the system to 'overbuild' wind, which reduces the need for flexible capacity. In essence, the model builds sufficient wind capacity that even during hours with low capacity factors, the wind can serve much of the load. Some natural gas-fired generation and energy storage are used to supplement wind production in these cases.

Overall, the representative days that are selected by the three clustering techniques result in investment levels that follow those given by the unclustered data. We can quantify the extent to which the investments that are determined by using clustered data match those that are determined by using unclustered data, through the use of a normalized RMSD metric. We define the normalized RMSD in investments in

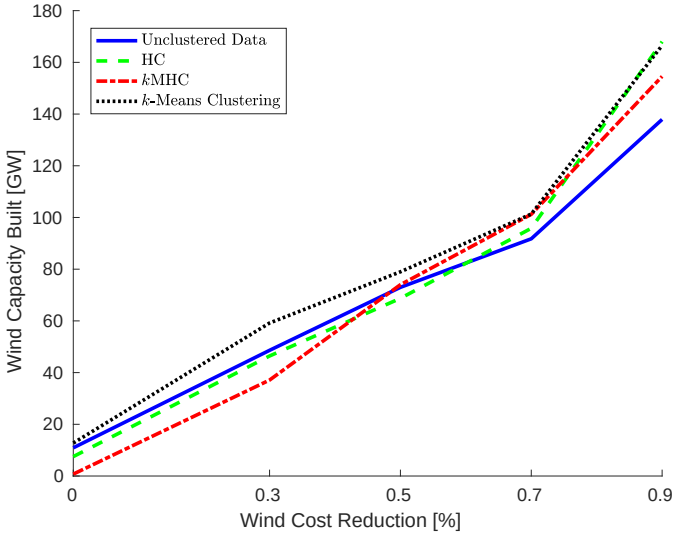

Fig. 5. Wind capacity that is built as a function of the reduction in wind-investment cost relative to baseline using unclustered and clustered data.

technology τ as:

$$\frac{\sqrt{\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\left(z^{C}_{\tau,\gamma}-z^{U}_{\tau,\gamma}\right)^{2}}}{\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}z^{U}_{\tau,\gamma}}. \tag{23}$$

Table VI reports the RMSDs for the five different technologies that are built (note that solar and nuclear are not built, because of their relatively high investment costs). The table shows that the HC technique results in investments that are closest to those that are given by the unclustered data. However, the HC and kMHC techniques both perform better than k-means clustering overall.

TABLE VI
NORMALIZED RMSD FOR INVESTMENTS GIVEN BY HC, kMHC, AND k-MEANS CLUSTERING

	HC	kMHC	k-Means Clustering
Coal	0.07	0.48	0.18
Natural Gas	0.12	0.18	0.20
Wind	0.19	0.14	0.20
Storage	0.19	0.22	0.32
Transmission	0.11	0.15	0.25

VI. CONCLUSIONS

This paper proposes two hierarchical clustering methods to select representative operating days for long-term capacity-expansion models. The use of representative operating periods reduces the computational cost of such models. Our methodology allows intertemporal operating constraints, such as storage state-of-charge-balance and generator ramping, to be captured in the investment model. At the same time, the clustering methods produce representative days that capture the important statistical features of operating-condition data. This includes temporal autocorrelations in the data and correlations between the locations that are modeled. This is primarly demonstrated by comparing the investment decisions that are made using the

clustered data to those made using the full unclustered dataset. We show that the HC and kMHC methods yield investment decisions that are comparable to those that are obtained using the full unclustered data.

Using a case study that is based on the Texan power system, we test our proposed methodologies and compare their performance to applying k-means clustering alone. We show that the representative days that are selected by our proposed methods capture the time dynamics in the load, solar, and wind data. We also demonstrate that they result in better overall investment decisions than applying k-means clustering alone does. We do this by examining investment levels with different wind-investment costs. Our own testing [18] also examines the impacts of changing other investment-model parameters (*e.g.*, solar-investment costs and operating costs of different technologies). This testing, which we do not present here for sake of brevity, demonstrate the same finding that the HC and kMHC methods yield investment decisions that are comparable to those that are obtained from using the full unclustered data. The kMHC technique reduces the computational burden of selecting representative days compared to HC, but this comes at some cost in terms of quality of the investment decisions.

The relatively good performance of the kMHC technique compared to HC method may seem surprising in light of the differences in the days that the two methods select (*cf.* Fig. 1). This finding highlights an important nuance in the selection of representative days. The goal is to find an ensemble of days that represent the different feature of load, wind, and solar patterns. Although there are some differences in the days that are selected by the HC and kMHC methods, they ultimately come up with ensembles with relatively similar patterns.

An important question underlying the use of our proposed clustering techniques is how many representative days should be selected. There is a clear tradeoff here. More representative days allows for higher fidelity in modeling operating conditions. Conversely, more days yields a larger and less tractable capacity-expansion model. The capacity-expansion model that we use in this paper is based on a more complex multistage, multiscale stochastic investment model [9]. Even with the use of the progressive hedging decomposition algorithm [28], that more complex model can take over a week to solve with 30 representative days if a sufficient number of scenarios and investment stages are included. Thus, judiciously selecting the minimum number of representative days possible is important from the perspective of maintaining a tractable capacity-expansion model. Indeed, one may argue that the shortcomings of applying k-means clustering alone can be overcome by use of more representative days (with less computational time involved than employing the HC or kMHC methods). This is a naïve view, however, because it neglects the cost of solving the resulting capacity-expansion model. This further demonstrates a contribution of our work, as the HC or kMHC methods do a better job than k-means clustering in judiciously selecting a small number of representative operating days for investment modeling.

Our analysis focuses on using clustering to obtain representative operating days. One could naturally use our proposed methods to select operating hours, for instance if it is not necessary to model intertemporal operating constraints. Conversely, the methods could also be used to select operating weeks if, for instance, interday energy storage is an important modeling feature. A benefit of our proposed clustering methods relative to others that are proposed in the literature (even if selecting representative operating hours) is that they do not arbitrarily assign equal weights to different seasons of periods of the year, as illustrated by Fig. 1.

ACKNOWLEDGMENT

Thank you to Armin Sorooshian, the editors, and five anonymous reviewers for helpful suggestions, comments, and conversations.

REFERENCES

[1] M. C. Caramanis, R. D. Tabors, K. S. Nochur, and F. C. Schweppe, "The Introduction of Non-Dispatchable Technologies as Decision Variables in Long-Term Generation Expansion Models," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, pp. 2658–2667, August 1982.

[2] W. Short, P. Sullivan, T. Mai, M. Mowers, C. Uriarte, N. Blair, D. Heimiller, and A. Martinez, "Regional Energy Deployment System (ReEDS)," National Renewable Energy Laboratory, Tech. Rep. NREL/TP-6A20-46534, December 2011.

[3] A. Pina, C. Silva, and P. Ferrão, "Modeling hourly electricity dynamics for policy making in long-term scenarios," *Energy Policy*, vol. 39, pp. 4692–4702, September 2011.

[4] L. Baringo and A. J. Conejo, "Correlated wind-power production and electric load scenarios for investment decisions," *Applied Energy*, vol. 101, pp. 475–482, January 2013.

[5] S. Wogrin, P. Dueñas, A. Delgadillo, and J. Reneses, "A New Approach to Model Load Levels in Electric Power Systems With High Renewable Penetration," *IEEE Transactions on Power Systems*, vol. 29, pp. 2210–2218, September 2014.

[6] K. Poncelet, H. Höschle, E. Delarue, A. Virag, and W. D'haeseleer, "Selecting Representative Days for Capturing the Implications of Integrating Intermittent Renewables in Generation Expansion Planning Problems," *IEEE Transactions on Power Systems*, vol. 32, pp. 1936–1948, May 2017.

[7] R. Alvarez, A. Moser, and C. A. Rahmann, "Novel Methodology for Selecting Representative Operating Points for the TNEP," *IEEE Transactions on Power Systems*, vol. 32, pp. 2234–2242, May 2017.

[8] Q. Ploussard, L. Olmos, and A. Ramos, "An Operational State Aggregation Technique for Transmission Expansion Planning Based on Line Benefits," *IEEE Transactions on Power Systems*, vol. 32, pp. 2744–2755, July 2017.

[9] Y. Liu, R. Sioshansi, and A. J. Conejo, "Multistage Stochastic Investment Planning with Multiscale Representation of Uncertainties and Decisions," *IEEE Transactions on Power Systems*, 2017, in press.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, pp. 264–323, September 1999.

[12] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River, New Jersey: Pearson Education, 2008.

[13] J. H. Ward, Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[14] S. I. Ao, K. Yip, M. Ng, D. Cheung, P.-Y. Fong, I. Melhado, and P. C. Sham, "CLUSTAG: Hierarchical Clustering and Graph Methods for Selecting Tag SNPs," *Bioinformatics*, vol. 21, pp. 1735–1736, 15 April 2005.

[15] J. Bien and R. Tibshirani, "Hierarchical Clustering With Prototypes via Minimax Linkage," *Journal of the American Statistical Association*, vol. 106, pp. 1075–1084, 2011.

[16] D. Gami, R. Sioshansi, and P. Denholm, "Data Challenges in Estimating the Capacity Value of Solar Photovoltaics," *IEEE Journal of Photovoltaics*, vol. 7, pp. 1065–1073, July 2017.

[17] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, February 1978.
[18] Y. Liu, "Electricity Capacity Investments and Cost Recovery with Renewables," Ph.D. dissertation, The Ohio State University, Columbus, Ohio, USA, August 2016.
[19] F. Graves, T. Jenkin, and D. Murphy, "Opportunities for Electricity Storage in Deregulating Markets," *The Electricity Journal*, vol. 12, pp. 46–56, October 1999.
[20] R. Sioshansi, P. Denholm, T. Jenkin, and J. Weiss, "Estimating the Value of Electricity Storage in PJM: Arbitrage and Some Welfare Effects," *Energy Economics*, vol. 31, pp. 269–277, March 2009.
[21] Y. Liu, M. C. Roberts, and R. Sioshansi, "A Vector Autoregression Weather Model for Electricity Supply and Demand Modeling," *Journal of Modern Power Systems and Clean Energy*, 2017, in press.
[22] A. Pielow, R. Sioshansi, and M. C. Roberts, "Modeling Short-run Electricity Demand with Long-term Growth Rates and Consumer Price Elasticity in Commercial and Industrial Sectors," *Energy*, vol. 46, pp. 533–540, October 2012.
[23] M. Muratori, M. C. Roberts, R. Sioshansi, V. Marano, and G. Rizzoni, "A highly resolved modeling technique to simulate residential power demand," *Applied Energy*, vol. 107, pp. 465–473, July 2013.
[24] *Annual Energy Outlook 2014*, DOE/EIA-0383 (2014) ed., U.S. Energy Information Administration, April 2014.
[25] "Cost and Performance Data for Power Generation Technologies," Tech. Rep., February 2012, prepared for the National Renewable Energy Laboratory.
[26] T. Mai, D. Sandor, R. Wiser, and T. R. Schneider, "Renewable Electricity Futures Study: Executive Summary," National Renewable Energy Laboratory, Golden, Colorado, Tech. Rep. NREL/TP-6A20-52409-ES, 2012.
[27] R. Masiello, K. Vu, L. Deng, A. Abrams, K. Corfee, J. Harrison, D. Hawkins, and K. Yagnik, "Research Evaluation of Wind Generation, Solar Generation, and Storage Impact on the California Grid," California Energy Commission, Sacramento, California, Tech. Rep. CEC-500-2010-010, June 2010.
[28] R. T. Rockafellar and R. J.-B. Wets, "Scenario and policy aggregation in optimization under uncertainty," *Mathematics of Operations Research*, vol. 16, pp. 119–147, November 1991.

Yixian Liu holds the B.E. degree in logistics engineering from Tianjin University and the M.S. and Ph.D. degrees in industrial and systems engineering from The Ohio State University.

Her research interests focus on weather forecasting, electricity capacity investment, and energy-policy analysis.

Ramteen Sioshansi (M'11–SM'12) holds the B.A. degree in economics and applied mathematics and the M.S. and Ph.D. degrees in industrial engineering and operations research from the University of California, Berkeley, and an M.Sc. in econometrics and mathematical economics from The London School of Economics and Political Science.

He is an associate professor in the Department of Integrated Systems Engineering at The Ohio State University, Columbus, OH. His research focuses on renewable and sustainable energy system analysis and the design of restructured competitive electricity markets.

Antonio J. Conejo (F'04) received the M.S. degree from the Massachusetts Institute of Technology, Cambridge, MA, in 1987, and the Ph.D. degree from the Royal Institute of Technology, Stockholm, Sweden, in 1990.

He is currently a professor in the Department of Integrated Systems Engineering and the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH. His research interests include control, operations, planning, economics and regulation of electric energy systems, as well as statistics and optimization theory and their applications.