The Case for a Simple Two-Sided Electricity Market

Alberto J. Lamadrid  
Lehigh University  
Bethlehem, PA, 18015  
Email: ajlamadrid@lehigh.edu

Wooyoung Jeon  
Chonnam National University  
Gwangju, South Korea  
Email: wyjeon@jnu.ac.kr

Hao Lu  
Australian Energy Market Operator (AEMO)  
Melbourne, Australia  
Email: hao.lu@aemo.com.au

Tim Mount  
Cornell University  
Ithaca, NY, 14853  
Email: tdm2@cornell.edu

Abstract—This paper builds on the results from our earlier research on the design of electricity markets that have to accommodate the uncertainty associated with high penetrations of renewable sources of energy. The key results show that 1) distributed storage (deferrable demand) is an effective way to reduce total system costs, 2) a simple market structure for energy allows aggregators to meet their customers’ energy needs and provide ramping services to the system operator, and 3) using a receding-horizon optimization to dispatch units for the next market time-step benefits from the availability of more accurate forecasts of renewable generation and allows market participants to adjust their bids and offers in response to this new information. In our two-sided market, distributed storage in the form of deferrable demand is controlled locally by independent aggregators to minimize their expected payments for energy in the wholesale market, subject to meeting the energy needs of their customers. In addition, these aggregators are responsible for maintaining a stable power factor by installing local capabilities that automatically deal with local power imbalances. Failure to do this triggers penalties paid to the system operator.

Our earlier results have shown that it is optimal for an aggregator to submit demand bids into a day-ahead market that include threshold prices for charging and discharging storage and also ensure that the expected amount of stored energy is consistent with the capacity limits of their storage. Because departures from the expected daily pattern of renewable generation are generally persistent (highly positive serial correlated), it is likely that the system operator determines an optimum pattern of demand for the aggregator that violates the capacity limits of storage by the end of the 24-hour period. If the market uses a receding horizon, the results in this paper show that aggregators can modify their bids to ensure that the capacity limits of storage are never violated in the next market time-step.

In an empirical application, a stochastic form of multi-period security constrained unit commitment with optimal power flow (the MATPOWER Optimal Scheduling Tool, MOST) using a receding-horizon optimization determines the optimum dispatch and reserves for the next hour and forecasts of the nodal prices for the next 24 hours. The results show that locally controlled deferrable demand is almost as effective as centrally controlled deferrable demand as a way to reduce system costs and mitigate the variability of renewable generation. The additional advantage from using a receding horizon is that the system operator always charges/discharges the storage managed locally by aggregators within the capacity constraints of the storage.

I. INTRODUCTION

System Operators (SOs) of the bulk electricity system are dealing with increasing amounts of uncertainty and variability from both supply and demand resources. Wind and solar power are responsible for much of this problem. Even though a large effort is underway by both academic and industry researchers to develop stochastic models for managing power systems, there is still a significant need for tools that can reflect the uncertainty of renewable energy and generate the market signals needed in deregulated markets (e.g., the prices for energy and ancillary services). The recent literature related to the modeling of the SO problem with uncertain and variable resources can be broadly categorized into three approaches: stochastic programming, probabilistic optimization and robust optimization. Stochastic programming has dimensionality issues (e.g., [1], [2]) which are handled using sampling methods to select a subset of possible scenarios by, for example, focusing on the most influential ones [3]. Probabilistic optimization allows violation of some specified network constraints within a threshold [4], [5]. Robust optimization considers a lower bound on the total social benefits by looking at the worst case realizations before the actual system state is realized [6]. This paper uses a hybrid method, between stochastic programming and robust optimization, to model uncertainty and determine the optimum commitment and dispatch of the generating units and the corresponding total social benefits and costs incurred by the market participants. An important feature of this modeling framework is that the optimum amounts of generating capacity needed for both dispatching units and procuring ancillary services are determined endogenously for a 24-hour horizon. This framework makes it feasible to determine an optimum strategy for managing storage capacity to shift load from peak to off-peak hours as well as to provide ramping services.

The objective of the paper is to demonstrate the potential benefits of using a simple two-sided market for electric energy with a receding-horizon optimization in which aggregators control distributed storage, in the form of deferrable demand, to minimize the expected cost of purchasing energy from the grid to meet the energy needs of their customers. In the application, a market solution is determined each hour for the next 24 hours, but the actual dispatch is limited to the first hour of the horizon. In this way, it becomes practical to use more accurate, updated forecasts of variable resources, like wind generation, and by doing so, reduce the range of likely values for these variable resources. As a result, the amount of reserve capacity is lower and the required amount of capacity to meet the energy needs of their customers. In the application, a market solution is determined each hour for the next 24 hours, but the actual dispatch is limited to the first hour of the horizon. In this way, it becomes practical to use more accurate, updated forecasts of variable resources, like wind generation, and by doing so, reduce the range of likely values for these variable resources. As a result, the amount of reserve capacity is lower and the required amount of capacity to meet the energy needs of their customers. In the application, a market solution is determined each hour for the next 24 hours, but the actual dispatch is limited to the first hour of the horizon. In this way, it becomes practical to use more accurate, updated forecasts of variable resources, like wind generation, and by doing so, reduce the range of likely values for these variable resources.
generating capacity needed for ramping and the associated operating costs are lower. Given forecasts of nodal prices for the next 24 hours, aggregators submit bids to buy with price thresholds for charging and discharging storage capacity. These bids can be modified each hour to ensure that the limits of storage capacity are not surpassed for the next hour. In contrast, in a day-ahead market in which aggregators submit only one set of bids into the market, it is highly likely that the charging/discharging profiles selected by the SO violates the capacity bounds somewhere in the 24-hour horizon.

In a recent paper [7], we presented an econometric model for simulating hourly levels of wind speed at different locations. The simulated values from this model were then transformed to the corresponding levels of potential wind generation using a deterministic power curve representing a wind farm. The econometric model has an ARMAX structure to capture the highly positive autocorrelations of the residuals as well as the spatial correlations. For a typical location, the full model can explain roughly 80% of the total variability for one-hour ahead forecasts, but the least-squares fit, ignoring the autocorrelated structure of the residuals, can only explain 10% of the variability. The explanatory power of the forecasts deteriorates quickly as the number of steps ahead increases, and the explanatory power of forecasts made more than eight hours ahead falls to the least-squares fit of 10%. The implication is that the forecasts of wind generation used in a day-ahead market are relatively inaccurate for most of the hours considered in the optimization. In this earlier paper, a comparison of the optimum amount of ramping capacity committed using a forecast made one hour ahead versus a forecast made six hours ahead showed that the range of ramping needed with the one-hour ahead forecast was only one eighth of the range needed with the six-hour ahead forecast. Hence, the evidence from these results suggests that the potential cost savings from lower ramping needs using updated forecasts with a receding-horizon optimization could be substantial. In addition, fuel costs were also reduced because more of the potential wind generation was dispatched, and less spilled, using the updated forecasts.

This paper is structured as follows. Section II provides a brief overview of the distinctive features of our analytical framework, the context of the optimal dispatch problem and a description of the model used and the advantages that distinguish it from other models. Section III-B summarizes the results that we presented in [8] and shows how an aggregator can use stochastic price forecasts to determine the hourly price thresholds for charging and discharging and the corresponding bids to meet customers demand for energy services and minimize the expected cost of purchases of energy from the grid. Section IV builds on our previous work [9] [10] by comparing the characteristics of the stochastic inputs used with a day-ahead market with the updated inputs used with a receding horizon. An empirical example, based on a reduction of the network in New York and New England, demonstrates the sensitivity of the results to the accuracy of the stochastic inputs and shows that there are attainable savings in the cost of reserve capacity using a receding horizon. The results also show that the bids submitted by aggregators, with price thresholds for charging and discharging storage, provide an effective mechanism for approximating how the same storage would be charged/discharged optimally if it was controlled centrally by an SO. Furthermore, using a receding horizon with aggregators managing the storage ensures that the capacity constraints of storage are not violated. This corrects an inevitable problem with aggregators submitting bids into a fixed-horizon market because it is highly likely that the optimum plan determined by the SO violates the capacity constraints at some point during the 24-hour horizon. Some concluding remarks and suggestions for ongoing research are presented in Section V.

II. FORMULATION OF THE MODEL

The analytical framework used, the Matpower Optimal Scheduling Tool, MOST, can be characterized as a probabilistic hybrid of a stochastic program and a robust optimization that represents a stochastic form of Security Constrained Unit Commitment (SCUC). It determines the optimum levels of ancillary services and the Optimal Power Flow (OPF) endogenously for a 24-hour planning horizon, subject to network constraints [11].

This work draws on the stochastic optimization literature, [12], [13], [14], [15], [16], [17], [6] with an emphasis on the determination of the costs incurred by all participants in the system. The main differences between our approach and other approaches can be summarized by the following four points (i) energy and two kinds of ancillary services (contingency and load-following reserves) are co-optimized by solving for the optimal amounts of reserves endogenously as part of the solution set [18], (ii) the cost of ramping delivered is internalized by assigning a wear-and-tear cost to changes in the dispatch of individual generating units [19], (iii) demand and supply costs are treated symmetrically by pricing Load Not Served (LNS) at the Value Of Lost Load (VOLL), and (iv) Energy Storage Systems (ESS) are modeled as a special set of generators with transversality conditions that value the end-of-horizon states and allow ESS to provide both load shifting and ramping services. These characteristics are especially important when the uncertainty of the stochastic sources of generation is large [20]. Basically, greater uncertainty leads to higher costs with more generating capacity being committed to maintain system reliability.

The objective function for the SO is the probability-weighted total welfare of all of the participants in all system states, and it has the following seven components (a) The cost of energy delivered, (b) The cost of re-dispatching the system (e.g., deviations from contracts), (c) The benefit that consumers receive by having their load serviced (i.e., by avoiding the cost of shedding load), (d) The cost of reserves (up and down) for low probability events (e.g., to cover contingencies), (e) The cost of reserves for high probability events (e.g., to mitigate the variability of wind generation and load), (f) The cost incurred in the transitions to new system.
states (e.g., the wear-and-tear costs), and (g) the cost of stored energy at the end of the planning horizon. The stochastic resources (e.g., potential wind generation) are modeled as Markovian processes with a discrete probability distribution over a finite number of states for each time period (e.g., five for each hour in the analysis).

A simplified objective function is presented in (2).

$$\min f(x) = f_P(p, p_r, p_m) + f_v(r_s, r_+, r_-) + f_x(p) + f_u(\delta, \delta_m) + f_{uc}(v, w)$$

where,

$$f_P(p, p_r, p_m) = \sum_{t \in T} \sum_{j \in J} \sum_{k \in K^c_j} \psi_{tjk}^c \left[C^c(p_{tjk}) + C^c_R(p_{tjk}) + C^c_{-R}(p_{tjk})\right]$$

$$f_v(r_s, r_+, r_-) = \sum_{t \in T} \sum_{j \in T} \left[C^v_R(r_{s+}) + C^v_R(r_{s-})\right]$$

$$f_x(p) = \sum_{t \in T} \sum_{j \in T} \sum_{j' \in T} \left[C^x_{j+}(p_{j'j}) + C^x_{j-}(p_{j'j})\right]$$

$$f_u(\delta, \delta_m) = \sum_{t \in T} \sum_{i \in I} \left[C^u_{\delta+}(\delta^i) + C^u_{\delta-}(\delta^i)\right]$$

$$f_{uc}(v, w) = \sum_{t \in T} \sum_{i \in I} \left[C^u_{v+}(0)w^i + C^u_{v+}(t^i) + C^u_{w+}(w^i)\right]$$

(2)

Table I summarizes the indices, functions and parameters considered in this reduced form, and Table II includes the optimization variables.

The constraints for the problem can be grouped into the following seven categories of components, (1) the full set of equality constraints (e.g., power balance equations), (2) the full set of inequality constraints (e.g., generator’s capability curves), (3) the set of constraints for reserve, redispatch and contract deviations, (4) the ramping limits for low probability events, (5) the ramping limits for high probability events, (6) the minimum startup and shutdown times, and (7) integrality constraints.

A central issue for using storage efficiently is to determine the optimum balance between shifting load from high-price periods to low-price periods and providing ramping services to mitigate the hour-to-hour variability of generation from renewable sources. With stochastic inputs, this issue has important implications for how energy constraints in the model are imposed on storage capacity that are illustrated in Figure 1. For each hour, the amounts of energy charged/discharged from storage is typically different in the intact system states.
representing different levels of the stochastic inputs (e.g., the amount of potential wind generation). Hence, the optimum level of energy stored in a given hour is also stochastic. If the sequence of “worst-case” states (i.e., the system states with the highest/lowest charging levels) are used to impose the capacity constraints on stored energy, the capacity limits are generally reached in a relatively few time steps (hours). This situation is represented by the black triangles in Figure 1.

Both the upper and lower capacity limits are reached at time $t=4$, and if these constraints are imposed, the range of feasible levels of stored energy at time $t$ is reduced to almost zero. Imposing the worst-case constraints on amount of stored energy for a 24-hour planning horizon implies that storage only provides ramping services for the last few time periods in the planning horizon. For all of the earlier time periods, the optimum charging/discharging strategy is deterministic and no ramping services are supported. This is an unreasonable outcome, particularly when a receding horizon is used to commit generating units. As a less restrictive alternative, the capacity constraints are imposed on the expected level of stored energy (represented by the blue diamonds in Figure 1).

This approach makes it feasible for storage to provide ramping services throughout the planning horizon.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>DEFINITION OF OPTIMIZATION VARIABLES, SIMPLIFIED FORMULATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^{t}_{i,j/k}$</td>
<td>Active injection for unit $i$ in post-contingency state $k$ of state $j$ at time $t$.</td>
</tr>
<tr>
<td>$p^{t}_{i}$</td>
<td>Active power contract quantity for unit $i$ at time $t$.</td>
</tr>
<tr>
<td>$p^{t}<em>{i,j/k}, p^{t}</em>{i,j/k}$</td>
<td>Upward/downward deviation from active power contract quantity for unit $i$ in post-contingency state $k$ of state $j$ at time $t$.</td>
</tr>
<tr>
<td>$r^{t}_{i}$</td>
<td>Zonal reserve quantity provided by unit $i$ at time $t$.</td>
</tr>
<tr>
<td>$r^{t}<em>{i}, r^{t}</em>{i}, r^{t}_{i}$</td>
<td>Zonal active reserve quantity provided by unit $i$ at time $t$.</td>
</tr>
<tr>
<td>$g^{t}<em>{i,j,k} p^{t}</em>{i}$</td>
<td>Voltage angles and active injections for power flow in post-contingency state $k$ of state $j$ at time $t$.</td>
</tr>
<tr>
<td>$u^{t}_{i}$</td>
<td>Binary commitment state for unit $i$ in period $t$, 1 if unit is on-line, 0 otherwise.</td>
</tr>
<tr>
<td>$w^{t}<em>{i}, w^{t}</em>{i}$</td>
<td>Binary startup and shutdown states for unit $i$ in period $t$, 1 if unit has a startup/shutdown event in period $t$, 0 otherwise.</td>
</tr>
<tr>
<td>$p^{i,j,k}<em>{c, d}, p^{i,j,k}</em>{d, c}$</td>
<td>Charge/discharge power injections of storage unit $i$ in post-contingency state $k$ of state $s$ at time $t$.</td>
</tr>
<tr>
<td>$w^{t}<em>{i}, w^{t}</em>{i}$</td>
<td>Binary startup and shutdown states for unit $i$ in period $t$, 1 if unit has a startup/shutdown event in period $t$, 0 otherwise.</td>
</tr>
</tbody>
</table>

### III. SPECIFYING INPUTS FOR THE MODEL

The specifications of the input data for MOST are based on publicly available sources and include the physical constraints of a test network and the installed generating units, the stochastic characteristics of potential wind generation and load, and the energy and power capacities of Deferrable Demand (DD). These inputs are essentially the same as the inputs described in [10], and this section is a summary of Section III of that paper.

The test network is a 36-bus reduction of a New York and New England centric version of the Northeast Power Coordinating Council (NPCC) network [21]. As accommodating the variability of generation from wind farms is an important feature of the analysis, the model includes the cost of using conventional generating units to provide ramping services. These costs are consistent with the information in [22] and [23], and they are specified by fuel type using quadratic cost functions. The values are relatively high for base load units and lower for peaking units.

The stochastic inputs for potential wind generation at 16 locations (nine in New York and seven in New England) and load for seven regions (four in New York and three in New England) are derived from estimated time-series models using hourly data for temperature, wind speed and load. These models provide the stochastic inputs for MOST.

The amounts of PWG dispatched and the load served are endogenous in MOST, and a typical optimal dispatch uses all PWG in the low wind states but may spill some PWG in the high wind states to reduce the need for ramping. In other words, even though wind generation is free, it may be more efficient to use conventional generation to avoid some ramping costs. In a similar way, some load may not be served, particularly in the rare contingency states. However, shedding load is expensive, and the specified VOLL is $10,000/MWh for urban areas and $5,000/MWh for rural areas.

The specifications of DD consider only thermal storage for space cooling in this paper because air conditioning is the main cause of the annual peak system load, and this peak determines the system requirements for adequacy. In other words, reducing the peak load reduces the amount of installed generating capacity needed to meet standards for generation adequacy. The optimal management of storage determines when to charge (usually at night) and when to discharge (usually during peak load periods) the storage. Some customers have thermal storage but most do not. The energy capacity of thermal storage is 17 GWh, corresponding to one half of the total daily amount of electricity used for space cooling that is potentially deferrable.\(^2\) The maximum rate of charging rate is 2 GW, and the maximum rate of discharging is almost 3 GW.

The technical characteristics of storage are based on the products described in the reports by Evapco [25] and Calmac [26]. The hourly ice building power rate is 12% and the hourly ice melting power rate is 16.7% of the total storage capacity, but these rates vary by the number of chillers installed. The specified round-trip efficiency of 86% is based on an average Energy Efficiency Ratio (EER) of 8.8 for thermal storage compared to an EER of 10.2 for a conventional air conditioner.

\(^2\)This is the same amount as our previous papers.
For each hour, the aggregate PWG across the 16 sites is ranked and then assigned to one of five bins, representing five different system states. This assignment is then used to compute the transition probabilities from each system state in one hour to each system state in the following hour. Finally, the same assignment of total PWG to bins is applied to the values of PWG for individual sites, and for each hour, the average values of PWG for the five bins at each site determine the amount of PWG for the five possible system states.

When a receding horizon is used, the wind inputs are updated each hour for the following 24 hours, and a new set of transition probabilities and state-level values of PWG are computed. It is assumed that the daily profile of load in the second day is identical to the profile in the first day. In other words, updating the profile of load corresponds to taking the first observation in the profile and making it the last observation for the next update. For PWG, it is assumed that the expected hourly values of wind speeds in step one (i.e. the daily profiles of expected wind speeds used with the fixed horizon) are equal to the realized values of wind speed that are used to update the forecasts of wind speeds at each step. The rationale for doing this is because the expected forecasted levels of PWG in a fixed-horizon optimization is likely to be different from the actual realized levels of PWG. Due to the fact that typical forecasting errors for wind are highly positively correlated, this implies that the actual realized levels have a tendency to be consistently above/below the forecasted levels. When this happens, the corresponding generation costs using a fixed horizon tend to be consistently higher/lower than they are using a receding horizon based on updated forecasts of wind because wind generation displaces more expensive conventional generation. As a result, it is difficult to draw general conclusions from a comparison of operating costs between the fixed and receding horizon for a particular realization of PWG. The realized levels of PWG in our analysis turned out to be substantially higher than the forecasted levels. For the empirical analysis, we adjusted the means of the forecasts for the fixed horizon and made them equal to the realized levels. Consequently, the comparisons between the fixed and receding horizons are more meaningful but also favor the fixed horizon approach. A more complete analysis evaluating the effects of many different realizations of PWG is left for future research. Nevertheless, using the adjusted forecasts of PWG for the fixed horizon leads to lower levels of wind generation than the receding horizon because more PWG is spilled when the uncertainty is larger.

B. Modeling the Behavior of an Aggregator

Given the complexity of managing a power system with an increasing number and variety of Distributed Energy Resources (DER), it is highly likely that the DER is, in the future, controlled by a Distributed System Operator (DSO). One possibility is that a DSO could implement instructions received from the SO, and in this situation, the DER is effectively centrally controlled by the SO. This type of management hierarchy is implicitly assumed in most of our previous research on distributed storage. An alternative organization is for the DER to be managed locally on behalf of the customers. In this situation, the DSO represents an aggregator who participates in the wholesale market by submitting hourly bids and offers. With the DD from thermal storage, the purchase of electric energy from the grid to charge the storage can be decoupled from delivering the required cooling services. Consequently, unlike, for example, the demand response from turning up thermostats when electricity is expensive, DD provides a non-disruptive way to modify the load profile of purchases from the grid. The downside is that more electric energy must be purchased to provide the same level of service with DD due to the round-trip inefficiency of the thermal storage.

We assume for this analysis that the objective of an aggregator is to minimize the expected cost of net-purchases from the grid and, at the same time, ensure that all energy services are delivered to customers when they are needed. Hence, the basic question is how should the aggregators structure their bids into the wholesale market to buy energy when it is less expensive and also provide ramping services. This latter activity is essential if the aggregator is going to be effective at mimicking the optimum management of DD by an SO. Initially, we anticipated that the aggregators would have to participate in both the energy market and some form of ramping market to get the incentives needed to provide ramping services. However, we show below that this is not the case, and a simple structure of bids in the energy market with price thresholds for charging and discharging DD storage is sufficient. The requirements are that the aggregators know the hourly levels of energy services needed and have access to stochastic forecasts of prices for the next 24 hours.\(^3\)

Define \(D_t\) as the amount of DD energy that needs to be delivered to customers in hour \(t\) and \(\text{prob}(P_t)\) as the corresponding probability density function for the wholesale price of energy. The hourly levels of DD can be met either by direct purchases of energy from the grid or by discharging storage. The aggregator’s objective for managing the storage capacity of DD is to minimize the expected cost of meeting the demand profile, \(D_t\), and it is shown in [27] that, for any

\[\text{Define } D_t \text{ as the amount of DD energy that needs to be delivered to customers in hour } t \text{ and } \text{prob}(P_t) \text{ as the corresponding probability density function for the wholesale price of energy. The hourly levels of DD can be met either by direct purchases of energy from the grid or by discharging storage. The aggregator’s objective for managing the storage capacity of DD is to minimize the expected cost of meeting the demand profile, } D_t, \text{ and it is shown in [27] that, for any}\]

\[^3\text{If the price forecasts are deterministic, the optimum hourly bids are a fixed profile of hourly purchases and no ramping services are provided.}\]

---

3. For each hour, the aggregate PWG across the 16 sites is ranked and then assigned to one of five bins, representing five different system states. This assignment is then used to compute the transition probabilities from each system state in one hour to each system state in the following hour. Finally, the same assignment of total PWG to bins is applied to the values of PWG for individual sites, and for each hour, the average values of PWG for the five bins at each site determine the amount of PWG for the five possible system states.

When a receding horizon is used, the wind inputs are updated each hour for the following 24 hours, and a new set of transition probabilities and state-level values of PWG are computed. It is assumed that the daily profile of load in the second day is identical to the profile in the first day. In other words, updating the profile of load corresponds to taking the first observation in the profile and making it the last observation for the next update. For PWG, it is assumed that the expected hourly values of wind speeds in step one (i.e. the daily profiles of expected wind speeds used with the fixed horizon) are equal to the realized values of wind speed that are used to update the forecasts of wind speeds at each step. The rationale for doing this is because the expected forecasted levels of PWG in a fixed-horizon optimization is likely to be different from the actual realized levels of PWG. Due to the fact that typical forecasting errors for wind are highly positively correlated, this implies that the actual realized levels have a tendency to be consistently above/below the forecasted levels. When this happens, the corresponding generation costs using a fixed horizon tend to be consistently higher/lower than they are using a receding horizon based on updated forecasts of wind because wind generation displaces more expensive conventional generation. As a result, it is difficult to draw general conclusions from a comparison of operating costs between the fixed and receding horizon for a particular realization of PWG. The realized levels of PWG in our analysis turned out to be substantially higher than the forecasted levels. For the empirical analysis, we adjusted the means of the forecasts for the fixed horizon and made them equal to the realized levels. Consequently, the comparisons between the fixed and receding horizons are more meaningful but also favor the fixed horizon approach. A more complete analysis evaluating the effects of many different realizations of PWG is left for future research. Nevertheless, using the adjusted forecasts of PWG for the fixed horizon leads to lower levels of wind generation than the receding horizon because more PWG is spilled when the uncertainty is larger.

B. Modeling the Behavior of an Aggregator

Given the complexity of managing a power system with an increasing number and variety of Distributed Energy Resources (DER), it is highly likely that the DER is, in the future, controlled by a Distributed System Operator (DSO). One possibility is that a DSO could implement instructions received from the SO, and in this situation, the DER is effectively centrally controlled by the SO. This type of management hierarchy is implicitly assumed in most of our previous research on distributed storage. An alternative organization is for the DER to be managed locally on behalf of the customers. In this situation, the DSO represents an aggregator who participates in the wholesale market by submitting hourly bids and offers. With the DD from thermal storage, the purchase of electric energy from the grid to charge the storage can be decoupled from delivering the required cooling services. Consequently, unlike, for example, the demand response from turning up thermostats when electricity is expensive, DD provides a non-disruptive way to modify the load profile of purchases from the grid. The downside is that more electric energy must be purchased to provide the same level of service with DD due to the round-trip inefficiency of the thermal storage.

We assume for this analysis that the objective of an aggregator is to minimize the expected cost of net-purchases from the grid and, at the same time, ensure that all energy services are delivered to customers when they are needed. Hence, the basic question is how should the aggregators structure their bids into the wholesale market to buy energy when it is less expensive and also provide ramping services. This latter activity is essential if the aggregator is going to be effective at mimicking the optimum management of DD by an SO. Initially, we anticipated that the aggregators would have to participate in both the energy market and some form of ramping market to get the incentives needed to provide ramping services. However, we show below that this is not the case, and a simple structure of bids in the energy market with price thresholds for charging and discharging DD storage is sufficient. The requirements are that the aggregators know the hourly levels of energy services needed and have access to stochastic forecasts of prices for the next 24 hours.\(^3\)

Define \(D_t\) as the amount of DD energy that needs to be delivered to customers in hour \(t\) and \(\text{prob}(P_t)\) as the corresponding probability density function for the wholesale price of energy. The hourly levels of DD can be met either by direct purchases of energy from the grid or by discharging storage. The aggregator’s objective for managing the storage capacity of DD is to minimize the expected cost of meeting the demand profile, \(D_t\), and it is shown in [27] that, for any

\[^3\text{If the price forecasts are deterministic, the optimum hourly bids are a fixed profile of hourly purchases and no ramping services are provided.}\]
given hour, this corresponds to charging storage, \( c \), when the realized price is less than a low threshold, \( L \), and discharging storage, \( d \), when the realized price is above a high threshold, \( H \). When \( L \leq P \leq H \), no charging or discharging occurs. The implied form of the demand curve for managing storage is illustrated in figure 2. Using this strategy, the actual purchase of energy from the grid is \((D_t + c - d) > 0\). For each hour, \( \text{Min}[d, D_t, S_{t-1}] \) is the upper limit on the amount discharged and \( \text{Min}[c, (S_{\text{max}} - S_{t-1})] \) is the upper limit on the amount charged, where \( S_t \) is the amount of energy stored at the end of period \( t \).

One of the complications of modeling storage that was discussed earlier in this section is how to impose the constraints on the accumulated amount of energy stored each hour to ensure it is non-negative and less than the energy capacity of the storage. Using the same rationale as before, the constraints are imposed on the expected amount of stored energy. In addition, it is assumed that the expected amount of stored energy at the end of the planning horizon is equal to the initial amount. Without a constraint of this type, the final amount of energy stored would always be zero and the initial amount would be treated as a free source of energy.

Finally, the round-trip inefficiency of storage implies that the amount of energy purchased from the grid is larger than the energy discharged to meet DD. The objective function for an aggregator can now be written as follows:

\[
\min_{L_t, H_t} \sum_{t \in T} \left( D_t E(P_t) + \sum_{i \in I} (c E(P_i | P_t \leq L_t) \text{prob}(P_t \leq L_t) - \text{Min}[d, D_t] E(P_i | P_t > H_t) \text{prob}(P_t > H_t)) \right),
\]

subject to:

\[
0 \leq S_0 + \sum_{t \in T} (\text{prob}(P_t \leq L_t) e - \text{Min}[d, D_t] \text{prob}(P_t > H_t)) \leq S_{\text{max}} \quad \forall t,
\]

where \( T = 24 \), \( e \) is the round-trip efficiency of storage, \( S_0 \) is the initial charge of the storage, and \( S_{\text{max}} \) is the energy capacity of storage.

The full set of first-order conditions for minimizing the expected cost of the energy purchased to meet the hourly levels of \( D_t \) are derived in [27]. However, the optimum strategy is very simple and intuitive if the expected amount of stored energy, \( E(S_t) \), never reaches the capacity limits for the energy stored. First, both the high and low price thresholds remain constant for the whole planning horizon, and second, the optimum low threshold, \( L^* \), is equal to the high threshold times the round-trip efficiency, \( H^* e \). Hence, the price arbitrage between the two thresholds must be large enough to compensate for the round-trip inefficiency of storage, and the optimum thresholds are chosen to make the expected total amount of energy discharged, when the price is above \( H^* \), equal to the expected total amount of energy charged, when the price is below \( L^* \), times the round-trip efficiency. In practice, it is quite possible that the storage is charged in one system state and discharged in another for the same hour, and this capability implies that DD can provide ramping services even though there is no formal market for ramping. Consequently, the benefits of the aggregator’s strategy for managing storage are not limited to minimizing the expected cost of meeting the DD requirements. The bid strategy also provides the flexibility needed to deliver ramping services to the SO even though the nodal price of energy is the only market signal.

In a power system with a high penetration of renewable generation, flexibility in demand is highly valuable because the amount of renewable generation available in any hour is uncertain. In general, for a system with no demand response, all potential renewable generation would be dispatched if the realized amount of generation is lower than expected. In addition, it may also be necessary to dispatch reserve units to avoid shedding load. When the potential generation is higher than expected, some generation may be spilled to avoid ramping down units that are already generating. Because the price of energy is typically high when the realized wind generation is lower than expected and low when it is higher than expected, a self-interested aggregator provides ramping services to the SO even though no explicit instructions to do this have been given. The two main objectives of the next section are to compare the performance of 1) the decentralized control of DD by aggregators with centrally controlled DD managed by an SO, and 2) using a fixed 24-hour horizon in a day-ahead market with a receding horizon in which inputs are updated each hour and the system is re-optimized.

C. The Four Cases Analyzed

The results presented in the next section summarize the costs of serving a given demand profile for a 24-hour period for four different cases on the annual peak load day using both a fixed and a receding horizon. The analysis assumes that the wholesale market is deregulated and run by an SO.

The following four cases are evaluated.

1) Case 1: Base case
2) Case 2: Case 1 + 16GW of New Wind Capacity at 16 locations
3) Case 3: Case 2 + 17GWh of DD Storage at 5 load centers managed by the SO
4) Case 4: Case 3 with DD Storage managed by Aggregators

IV. EMPIRICAL RESULTS

Table III summarizes the optimum levels of operation for the four cases using a fixed 24-hour horizon and the corresponding four cases using a receding-horizon based on 24 separate optimizations with updated forecasts of PWG for each hour of the day.

The four cases using the conventional fixed-horizon optimization in table III replicate the results presented in earlier papers. The extra wind capacity in Case 2 displaces over 15% of the conventional generation in Case 1, but at the same

\(^4\)An aggregator submitting a high threshold price for discharging and a low threshold price for charging reduces purchases from the grid if the price is high enough and increases purchases if the price is low enough.
TABLE III
SUMMARY OF THE OPTIMUM OPERATING LEVELS FOR THE FIXED AND RECEDING HORIZONS

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. E[Wind Generation] (MWh)</td>
<td>689</td>
<td>206,117</td>
<td>208,433</td>
<td>209,362</td>
</tr>
<tr>
<td>2. E[Conventional Generation] (MWh)</td>
<td>1,268,793</td>
<td>1,063,575</td>
<td>1,063,575</td>
<td>1,058,375</td>
</tr>
<tr>
<td>3. LF Up Reserve (MW)</td>
<td>22,030</td>
<td>35,049</td>
<td>25,084</td>
<td>25,756</td>
</tr>
<tr>
<td>4. LF Down Reserve (MW)</td>
<td>20,360</td>
<td>31,072</td>
<td>23,324</td>
<td>20,954</td>
</tr>
<tr>
<td>5. Contingency Reserve (MW)</td>
<td>18,087</td>
<td>25,038</td>
<td>10,309</td>
<td>10,785</td>
</tr>
<tr>
<td>Conventional Generation, Max Intact States (MW)</td>
<td>62,100</td>
<td>56,985</td>
<td>54,962</td>
<td>55,052</td>
</tr>
<tr>
<td>Conventional Generation, Max (MW)</td>
<td>63,078</td>
<td>57,857</td>
<td>55,842</td>
<td>55,913</td>
</tr>
</tbody>
</table>

Fig. 3. Expected Operating Costs for Case 2

Turning now to the differences in results between the fixed and receding horizons, the differences for Case 1 are trivial because the initial amount of wind capacity is very small, but in Case 2 with 16 GW of additional wind capacity, the receding horizon does lead to lower average costs than the fixed horizon, particularly in the early hours of the morning when most charging occurs (see figure 3). The main reasons, based on the results in table III, are that using the receding horizon leads to 1) slightly less of the PWG is spilled, 2) less up and down reserves for ramping are needed because the updated forecasts of PWG are more accurate, 3) less contingency reserves, and 4) a lower maximum commitment of conventional capacity. These positive effects are even larger in Case 3 with deferrable demand. However, the overall benefits of the receding horizon are quite modest compared to our initial expectations. A possible reason for this is that adjusting the mean hourly forecasts of PWG for the fixed horizon to make them equal to the hourly levels of PWG that are actually realized using the receding horizon definitely favors the fixed horizon. This conclusion of modest benefits using a receding-horizon optimization may well change if a new comparison is made using many different realizations of PWG in future evaluations.

Using a receding horizon, the results for the aggregators in
Case 4, shown in table III, are not as close to the results for the SO in Case 3 as they are using a fixed horizon.\(^7\) There is no obvious reason for the larger differences between Cases 3 and 4, but it is likely that the forecasted prices used by the aggregators to determine their bids are systematically higher/lower than the prices actually determined by the SO in Case 4. Efforts to characterize the differences between Cases 3 and 4 are continuing. Nevertheless, the simple strategy for energy bids used by aggregators still reduces the amount of reserve capacity needed from 46.8 GW/day in Case 2\(^8\), with additional wind capacity but no deferrable demand, to 35.4 GW/day in Case 4. However, this is still substantially more than the 30.3 GW/day needed in Case 3 when the SO manages the deferrable demand. Finally, in Cases 3 and 4 with deferrable demand, the reduction in the total capacity needed for adequacy compared to Case 2 is not as large with a receding horizon as it is with a fixed horizon even though the reserve capacity needed is lower. This is another issue that is still being investigated, but it suggests that using a receding horizon may put relatively more weight on providing ramping services than shifting load compared to using a fixed horizon. This issue is important because reducing the capacity needed for system adequacy is an effective way to reduce total system costs.

An important feature of Case 4 using a receding horizon is that the capacity limits on storage are never exceeded because the updating of the bids by the aggregators each hour sets feasible limits on the amounts of charging/discharging that are feasible in the next hour given the amounts of energy stored at the end of the previous hour. This feature is illustrated in figure 4, and the infeasible discharging that occurs using a fixed horizon probably biases the amounts of reserve capacity needed and operating costs downward.

Table IV summarizes the components of the daily operating costs for all of the cases shown in table III and it reinforces the results discussed above. The main savings in cost from Case 1 for both the fixed and receding horizons come from displacing fossil fuel generation by wind generation. The extra cost of ramping in Case 2 using a fixed horizon is small compared to the reduction in the cost of conventional generation. As a result, the value of reducing the cost of ramping by adding deferrable demand or updating forecasts of PWG has only a minor effect on total operating costs. The main result is that aggregators participating in only the energy market in Case 4 do as well as the SO in Case 3, and if a receding horizon is used, the actual charging/discharging determined by the SO in Case 4 is feasible and the capacity limits on storage are never exceeded.

\(^7\)In this analysis, the high and low price thresholds are held constant and not updated each hour. Procedures for updating the thresholds, and most importantly, dealing with the effects on bids when energy capacity limits are reached are still being developed.

\(^8\)The corresponding amount using a fixed horizon is 60.1 GW/day in Case 2, and the reduction using a receding horizon is indicative of the value of more accurate forecasts of PWG

V. SUMMARY AND CONCLUSIONS

This paper presents and evaluates a proposal for improving the operations of an electricity system when there is a high level of uncertainty about stochastic inputs by implementing a receding horizon that incorporates updated forecasts of these inputs, and in the empirical application, the stochastic inputs are the hourly amounts of potential wind generation at different locations. The analysis uses a stochastic form of multi-period Security Constrained Unit Commitment (SCUC), the Matpower Optimal Scheduling Tool MOST, and the forecasts of the potential wind generation at different locations are derived from estimated ARMAX models of wind speeds. The empirical application uses a reduced network representing New York State and New England on a hot summer day when the system peak load occurs. The three cases of most interest are Case 2 with 16 GW of installed wind capacity, Case 3 with the addition of 17 GWh of deferrable demand (thermal storage) managed by a system operator, and Case 4 with the deferrable demand managed by an aggregator who submits bids/offers into the energy market. For each case, the analysis compares the results using a single fixed 24-hour horizon with the results using a receding horizon that solves for a 24-hour horizon each hour using updated forecasts of potential wind generation.

The typical forecasting errors for wind are highly positively correlated and incorporating this error structure in the ARMAX models provides most of the explanatory power (70% of total variation for one-hour ahead forecasts) compared to the deterministic components (10% of total variation). In addition, the explanatory power from the estimated residual structure decreases to almost zero after seven or eight hours ahead, and as a result, the amount of uncertainty about the actual levels of wind is much larger for the latter part of the day using a fixed horizon compared to a receding horizon based on hourly updated forecasts of potential wind generation. At each step, the receding horizon dispatches generating units for the next hour only even though the optimization determines the pattern of dispatch for the next 24 hours. In effect, the implied patterns of dispatch for hours \(t + 2\) to \(t + 24\) determined at hour \(t\) are superseded when better forecasts of wind are available at hour \(t + 1\). One complication in comparing the two methods is that in practice the actual realized levels of wind may be consistently higher/lower than the forecasted levels used in the fixed horizon, and therefore, the corresponding operating costs tend to be consistently lower/higher using a receding horizon. For this reason, we scaled the mean hourly forecasts used in the fixed horizon to be equal to the actual levels, and consequently, the only difference in inputs for our analysis is that there is less uncertainty about potential wind generation using the receding horizon.

The results from the empirical comparison of the two optimization procedures are consistent with our prior expectations in terms of direction but the magnitudes are somewhat smaller than we expected. Using the fixed horizon, wind generation displaces conventional generation in Case 2 but this requires
more conventional capacity committed for ramping reserves. In addition, the total amount of conventional capacity required for operating reliability is smaller, and this implies that less installed capacity is needed to maintain generation adequacy. Adding deferrable demand in Case 3 and Case 4 reduces this installed capacity further by shifting load from peak to off-peak hours, and deferrable demand also provides a substantial amount of the ramping services needed. Comparing these results with the equivalent cases using a receding horizon shows, for Cases 2, 3 and 4 that 1) less of the potential wind generation is spilled with a receding horizon, 2) less conventional capacity is needed for ramping, and 3) less installed conventional capacity is required for adequacy. In these three ways, the receding horizon has positive system benefits. Although the magnitudes of these benefits are small in this application, this may be partly due to rescaling the initial 24-hour forecast of wind to favor the fixed horizon optimization. It would be interesting to see in future research how these results change when the two approaches are evaluated with different realizations of wind, and in particular, when the actual wind is lower than expected and conventional sources of generation have to be increased by drawing down reserves.

The differences in the results between Cases 3 and 4 are small using a fixed horizon and show that the simple two-sided market in Case 4 with local control of deferrable demand by aggregators performs almost as well as Case 3 when a system operator manages everything. Using a receding horizon, the differences between Cases 3 and 4 are somewhat larger and work is continuing to determine the specific causes. In Case 4, the aggregators submit bids to buy energy with high threshold prices for discharging storage and low threshold prices for charging storage. Given stochastic forecasts of future prices, this bid structure minimizes the expected cost of supplying the energy needs of their customers, and at the same time, it implicitly provides ramping services to the system operator. If a receding-horizon optimization is used, the aggregators can update their bids at each time-step to ensure that the energy capacity limits of the storage are respected, and this corrects a problem that occurs frequently using a fixed-horizon optimization. It should be noted that updating bids in response to new price forecasts using a receding horizon market is very similar to the how the Australian electricity market has been operating for almost two decades.

Currently, there are two competing proposals for managing Distributed Energy Resources (DER). The first is to extend the logic of nodal pricing from the high-voltage grid to distribution systems. The second, which we favor, is to have DER managed locally by aggregators who submit bids into the energy market and work on behalf of their customers. We have demonstrated that this type of simple two-sided market with distributed storage controlled locally by aggregators can perform well. The market is simple because the aggregators only participate in the energy market and yet they still provide ramping capabilities to the system operator. The overall result is that the variability of renewable sources of energy can be accommodated. Negative prices when there is too much wind generation, for example, are a major incentive for charging storage. Similarly, high prices trigger a reduction in purchases from the grid. Ideally, distributed storage could smooth out the generation from conventional sources and increase their average capacity factors. This is a valuable improvement for the supply system because earnings in the wholesale market tend to fall for conventional generators when there are high penetrations of renewable generation.

Our empirical analysis assumes implicitly that the optimum results in all system states are in equilibrium with a unit power factor. However, with rooftop solar, for example, there are several local voltage problems when clouds pass overhead. We assume that these problems are managed locally through the installation of equipment such as smart inverters, because the response times needed to deal with voltage problems are typically shorter than the market time-step. A simple market mechanism that already exists for some wholesale customers is to provide incentives for aggregators to maintain a stable power favor by charging a penalty for violations. This combination of aggregators submitting bids for distributed storage in the energy market and equipment to manage voltage problems automatically is a practical, and possibly cost effective, solution that is consistent with the concept of “grid edge intelligence.”

ACKNOWLEDGMENTS

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000779, the National Science Foundation through the CRISP Type 2

---

### TABLE IV

<table>
<thead>
<tr>
<th>Composition of Wholesale Costs ($1000/day)</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E[Generation Cost]</td>
<td>30,947</td>
<td>22,871</td>
<td>19,565</td>
<td>19,575</td>
<td>31,992</td>
<td>21,845</td>
<td>19,324</td>
<td>19,559</td>
</tr>
<tr>
<td>E[Ramp Wear Cost]</td>
<td>2</td>
<td>198</td>
<td>28</td>
<td>29</td>
<td>5</td>
<td>29</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>LF Ramp-Up Reserve Cost</td>
<td>234</td>
<td>1,161</td>
<td>363</td>
<td>376</td>
<td>272</td>
<td>690</td>
<td>294</td>
<td>327</td>
</tr>
<tr>
<td>LF Ramp-Down Reserve Cost</td>
<td>204</td>
<td>387</td>
<td>239</td>
<td>219</td>
<td>244</td>
<td>304</td>
<td>157</td>
<td>106</td>
</tr>
<tr>
<td>Contingency Reserve Cost</td>
<td>88</td>
<td>122</td>
<td>50</td>
<td>53</td>
<td>96</td>
<td>92</td>
<td>36</td>
<td>54</td>
</tr>
<tr>
<td>E[Total Operating Cost]</td>
<td>31,475</td>
<td>24,739</td>
<td>21,320</td>
<td>20,252</td>
<td>32,607</td>
<td>22,960</td>
<td>21,505</td>
<td>20,131</td>
</tr>
</tbody>
</table>

---

8See, for example, the presentation at the National Science Foundation through the CRISP Type 2
grant #1541177 and the CyberSEES grant #1442858, the US Department of Energy through the Consortium for Electric Reliability Technology Solutions (CERTS). The authors would like to thank Ray D. Zimmerman, Carlos E. Murillo-Sanchez, Luis F. Zuluaga, and Robert J. Thomas for their comments and input. The authors are responsible for all conclusions presented in the paper and the views expressed have not been endorsed by the sponsoring agencies.

REFERENCES


