Is There Still Merit In The Merit Order Stack? The Impact of Dynamic Constraints on Optimal Plant Mix

Iain Staffell and Richard Green

Imperial College Business School, Imperial College London
i.staffell@imperial.ac.uk r.green@imperial.ac.uk

Originally published in IEEE Transactions on Power Systems
http://dx.doi.org/10.1109/TPWRS.2015.2407613

Abstract

The merit order stack is used to tackle a wide variety of problems involving electricity dispatch. The simplification it relies on is to neglect dynamic issues such as the cost of starting stations. This leads the merit order stack to give a poor representation of the hourly pattern of prices and under-estimate the optimal level of investment in both peaking and inflexible baseload generators, and thus their run-times by up to 30%.

We describe a simple method for incorporating start-up costs using a single equation derived from the load curve and station costs. The technique is demonstrated on the British electricity system in 2010 to test its performance against actual outturn, and in a 2020 scenario with increased wind capacity where it is compared to a dynamic unit-commitment scheduler. Our modification yields a better representation of electricity prices, and reduces the errors in capacity investment by a factor of two.

1 Introduction

We present a new heuristic that adds start-up costs to the merit order stack model of generation investment and operation decisions. Start-up costs can have a noticeable impact on electricity wholesale prices, and change the relationship between each unit's operating hours and total costs. Ignoring them means that a merit order stack optimization is likely to choose a plant mix that would be sub-optimal, had these starts been properly taken into account. We compare results from the traditional merit order stack approach with those from our heuristic, benchmarking them against historic output and price data, and the optimal capacity mix determined by a dynamic dispatch model that includes start-up costs and minimum load constraints. We show that the errors from the stack approach can be important, and that our heuristic improves the representation of prices by a factor of 1.6 (comparing to historic data), and reduces deviation from the optimal capacity mix by half (comparing to results from a more complex unit commitment model).

The need to decarbonize the power sector poses technical and economic challenges that need to be modelled in an appropriate manner. Many low-carbon generators are inflexible or intermittent, and the remaining stations will need to change their operating regimes [1]. Energy storage may mitigate this, but detailed technical modelling of generators and the grid is required to assess the true costs and benefits, and the best way to deploy storage devices [2]. The optimal capacity mix for generation is likely to change [3], and modelling is needed to explore this.

Investment decisions in power stations should not be assessed against a single scenario, however, but should consider a range of possible futures – for fuel prices, demand and the output of intermittent generators. This requires a model that can be solved rapidly while still producing sufficiently accurate results for each scenario, and thus a simplified approach. Our heuristic is simple enough for repeated simulation and more accurate than the traditional approach. Each user will have to make their own decision on whether it is accurate enough for them, depending on the use to which their model results are put.

We test our heuristic against the two more traditional approaches (simple merit order stack and full dynamic dispatch) with a model of the 2020 GB electricity system. This is characterized by having 30 GW of installed wind capacity (more than half the winter peak demand, and 1.5 times the system minimum demand), with relatively little interconnection or flexible hydro capacity. Using these models, we demonstrate how operating constraints change the dispatch of different stations, and the extent to which this would change the optimal plant mix.

The next section summarizes the strengths and weaknesses of the merit order stack and previous efforts to incorporate dynamic constraints into the approach. Sections 3 and 4 introduce our heuristic approach to start-up costs and to hydro scheduling, and the dynamic unit commitment model that we test it against. The two scenarios used to validate the heuristic are given in Section 5, and the main results in Section 6. The paper ends with brief conclusions on the performance and applicability of this technique.

2 Background

The merit order stack is a common approach to the problem of minimizing the sum of generators' investment and operating costs. It is equivalent to a linear program in which each station's operating costs are strictly proportional to its output and the only constraint is that output levels are less than that station's available capacity in each hour [4, 5]. Stations with low variable costs will always be dispatched at full capacity before any stations with higher variable costs are called upon, and so each station's output and operating hours will be approximately proportional to each other.

The station's total costs over the year are thus a linear function of its total output, often plotted as a screening curve which relates costs to operating hours per year. The number of operating hours for which two plant types would have equal costs can easily be calculated as the intersection of their screening curves. The optimal capacity mix is one for which every station is running for a number of hours over which it has lower total costs than any alternative.

The price of electricity can be obtained as the variable cost of the marginal unit. When the system is short of capacity, a rationing price may be used to reduce demand, either derived from the prices paid to large customers for load management or using the Value of Lost Load. The price of electricity can be very sensitive to number of outages reducing available capacity, and the relative simplicity of the merit order stack means that [6] could run an hour-by-hour model of California during its electricity crisis 100 times with different random outages to capture the non-linear nature of this relationship.

Simple models can bring useful insights when a problem is first approached; [7] assesses the benefits of introducing real-time electricity pricing in California, calculating the reduction in overall capacity and the change in wholesale price patterns that this would bring about. Other papers have used similar models to study the impact of the large-scale introduction of renewable generators on investment in conventional plants [8, 9, 10].

As a field develops, increasingly complex models and solution techniques are proposed. These might focus on the interaction between generation and transmission [11, 12] or combine generation investment with stochastic scheduling and dynamic operating constraints [13, 14, 15]. A long-term investment equilibrium and a short-term dispatch that takes account of operational constraints are combined in [3] to calculate the marginal value of wind and solar power at different penetration levels for a scenario of California in 2030. Similarly, [2] calculates the value and optimal capacity of electricity storage for the UK in a low carbon 2030 scenario, using a simultaneous optimization of investment and operation decisions, subject to plant- and network-level constraints. The disadvantage of these more complex models is that their computational requirements make them poorly suited for large-scale repeated simulations.

The time required to run a full dispatch model on a year of data makes it less suitable for many applications, including those involving stochasticity or multi-decadal optimizations. One solution is to use the dispatch model on a representative sample of days; for example, [16] shows how clustering techniques can be used to create these.

Another strand of research aims to find techniques for representing operational constraints in a simplified manner, allowing faster computation without losing the key features of the problem [17,

18]. One approach is to cluster similar units together rather than treating them individually, thus reducing the number of integer variables in the unit commitment problem [19, 20]. Another option is to keep the merit order stack approach but to add heuristics that approximate important features. [21] considers the amount of up- and down-regulation capacity that is required in each hour as a function of the level of wind generation, after setting capacity levels on the basis of a screening curve analysis. [22] compares a full unit commitment model and the simplified approach of economic dispatch subject only to ramping constraints and find that both give similar results, while the simplified model solves substantially faster. A more pessimistic view is taken in [23], which makes a similar comparison when modelling three different regions; ERCOT in the US, Finland and Ireland. Errors come from ignoring start costs, (typically) over-estimating the number of starts made and miscalculating energy costs because the simpler model will identify the wrong mix of operating plants. The authors suggest that better methods are needed.

One method of incorporating start costs is proposed in [24], using a two-stage process. The hour-by-hour demand profile is used to calculate the annual output and the number of starts required for the generating unit at each loading point (the position a plant occupies within the stack, measured in GW). A nuclear unit with low variable costs will be at a low loading point and will have no starts, running continuously through the year. An open cycle gas turbine will be at a high loading point with few operating hours and a high ratio of starts to output. The heuristic avoids stations shutting down for short periods so long as the rest of the fleet is able to reduce their output sufficiently without forcing any units below their minimum stable generation levels. The second stage of the model then calculates the cost that each kind of candidate plant would have, were it to match the generation profile calculated for a given loading point, and then selects the cheapest.

The aim of this paper is to propose a simpler method of modelling start-up costs that does not require a full unit commitment model or a multi-stage algorithm, such as [24]. Our reasoning for this is to save on the time required to both implement and then solve these more complex models, whilst improving the accuracy of results from the existing approach.

3 Extending the Merit Order Stack

As our basis, we use a merit order stack model with price responsive demand, as described in [7, 25, 26]. Plants are dispatched in order of increasing variable cost, and the price of electricity is either set to the variable cost of the marginal generator, or to the level that would reduce demand to the point where it could be supplied by the economically available capacity. An arbitrary number of stacks can be used so that availability, fuel and carbon prices can be varied by season or month. We extend this model in three ways:

- The cost of plant starts is factored into the price of electricity using a single equation, altering the profitability of plants and thus the optimum capacity to invest in.
- Hydro is scheduled using a peak-shaving algorithm that is split into multiple tranches to better replicate historic output patterns; and
- Each technology can be given a must-run output requirement, below which the fleet output cannot fall.

The must-run constraint is applied to nuclear stations, which in the UK do not operate below 90% capacity (except during outages). To avoid breaching this constraint (when demand net of wind is

low) wind power was spilled and the market price was set to –£50/MWh, the foregone subsidy which wind stations require as compensation. This increases the total cost for nuclear as more wind must be forced to spill. We do not model the must-run constraints that prevent mid-merit stations shutting down overnight; this is done by [24] but adds complexity to their model. We show below that our simpler approach gives adequate estimates of many key variables. Future research should address the improved accuracy that would result from more sophisticated modelling of mid-merit shutdown decisions.

The model with these extensions is implemented in Microsoft Excel 2010, and is available to download from http://hdl.handle.net/10044/1/12715.

3.1 Scheduling Hydro

The traditional approach to scheduling hydro in a merit-order stack model is to pre-treat the demand data with a peak-shaving algorithm, as used in [27] to model California. This finds the loading point (the demand in GW) above which hydro stations begin operating so as to produce the desired energy output over a given time-frame. For pumped storage, this level is chosen each day on the assumption that reservoirs are replenished fully each night. Pumping consumption can be allocated using the reverse logic: finding the loading point below which water should be pumped. Run of river hydro will usually have a minimum flow constraint, and the relevant capacity (and energy) is allocated to every hour. The flow constraints and water availability will vary over the year, and we schedule over three-month periods based on historic rainfall trends. When hydro is scheduled to run, net demand is reduced towards the loading point by as much as the hydro capacity allows.

Figure 1 demonstrates this algorithm in action: the first day has a short-lived peak with hours in which the full power capacity (measured in GW) of the pumped storage plants is used; the second day has a flatter peak and the full energy capacity (GWh) can be used without running at full power. The pumping of water is seen overnight, raising the system minimum demand. Run-of-river hydro begins to operate at the same loading point during both days, as the energy constraint is not applied to each day separately.

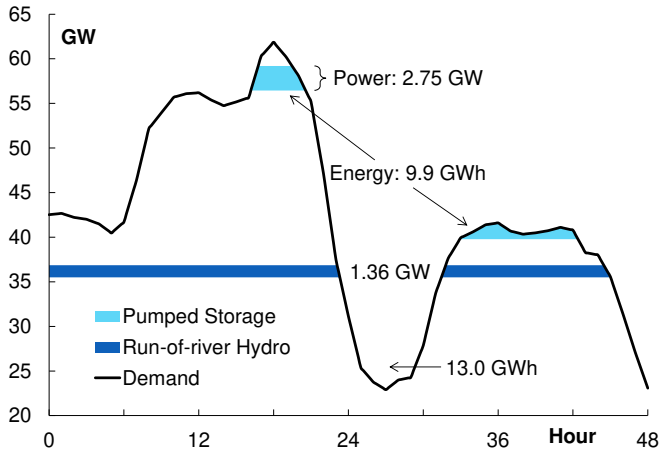

Fig. 1: Demonstration of the heuristic for allocating hydro resource.

In countries with little hydro capacity relative to demand, this algorithm assumes that the hydro fleet runs predominantly at either zero or full power, as demand is either below the loading point or above that level plus installed capacity. Taking the example in Figure 1 (which is based on the current GB

system), demand rarely lies within the window of 35.5–36.9 GW where run-of-river hydro should operate, and so the capacity factor is either 1 or 0 in such a ratio as to achieve the desired annual average. This occurs because simple peak shaving algorithms neglect the trade-off between operating in the energy and reserve markets. In reality, British run of river hydro spends only 300 hours a year operating above 80% of capacity, as it is more profitable to retain availability for balancing and reserve services.

This flaw is demonstrated in Figure 2, which shows the historic output-duration curves for British hydro stations (faint blue lines) against output from the basic peak shaving algorithm (dotted lines).
A solution is to split the hydro capacity into several tranches with different utilization rates. These are then allocated separately and their output summed. If statistics are available, this can be done on a per-plant basis, splitting each station into a must-run component to account for minimum flow constraints and a dispatchable portion. Alternatively, if time-series data of national hydro output are available, this can be segmented using techniques such as k-means clustering, as shown by the thin solid line in Figure 2, which is fitted to the historical data.

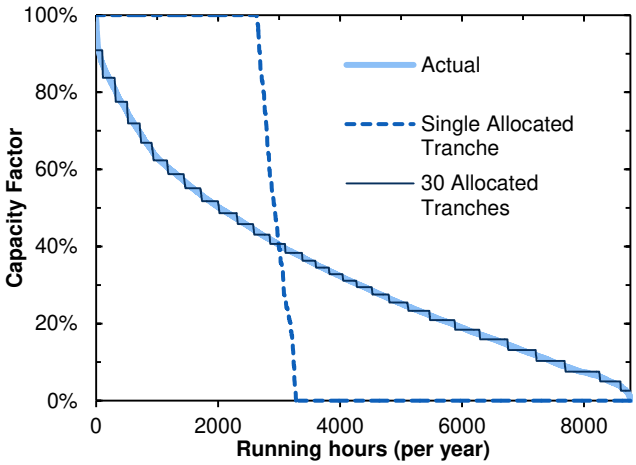

Fig. 2: Output-duration curves for British hydro in 2009–10, against the allocations made by the traditional application of the peak-shaving algorithm and the algorithm with multiple tranches fitted to the actual curve.

3.2 Plant Start Decisions

We incorporate start costs into the stack model by post-processing the price of electricity or the screening curves for each station, depending on the application. We therefore take account of start costs when calculating market prices and the optimal capacity mix, but not when calculating the hour-by-hour dispatch of stations. To the best of our knowledge, we are the first to propose such a technique.

As shown in Section 6.2, adding start costs has a noticeable effect on the accuracy of price simulations, validated against British data from 2009-10. They also change the optimal capacity mix, bringing it closer to that from a dynamic unit commitment model for simulations of 2020. We find that start costs do not have a large effect on our predictions for annual output from each plant type. There will be some periods in which a station is kept running at a reduced load in order to avoid starting again later, which increases its actual output compared to a stack model. However, it could be expected that these are roughly balanced by periods in which the station has to reduce output in

order to allow another station (with higher variable costs) to avoid a start. For applications concentrating on prices and annual profits, this simplification proves accurate enough.

We take a simple approach to estimate the number of start-ups that each plant will undergo: we start by assuming that any time national load (net of wind and hydro) rises above a given value, Q, the plant that sits at loading point Q in the stack plant must start up. Figure 3 demonstrates this, showing how many times plants at three loading points would start during a week. In this example, the number of starts per year is highest for mid-merit plant which must run for around 3,000 or 5,500 hours per year. This uses Billinton and Allan's method of modelling load transitions for the calculation of frequency duration indices [28]; the innovation here is to use these statistics to calculate plant starts.

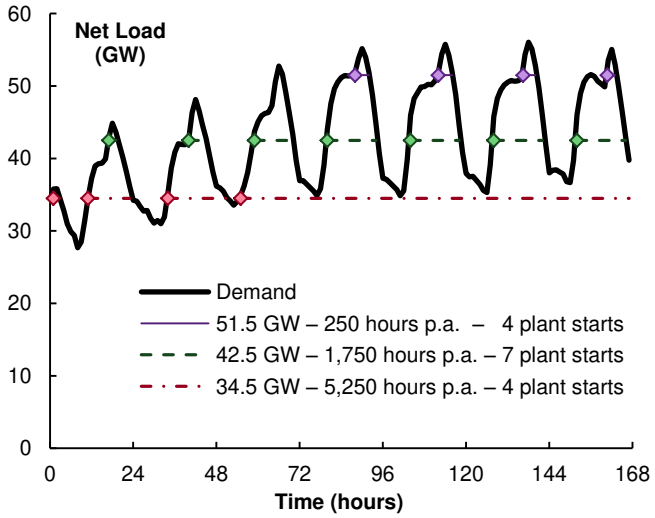

Fig. 3: Demonstrating a simple algorithm for calculating the number of start-ups for different levels of plant.

The profile of start-ups against loading point will change if different countries are considered, and over time as the national level of demand changes. We find that the relationship between the number of starts and the number of running hours is more consistent over time and between countries and use this in our heuristic, in contrast to [24]. The load duration curve can be used to map the loading point of a plant (in GW) to the number of hours that plant is required for, just as when constructing a screening curve. Figure 4 demonstrates this transformation, showing the similarity of the start-up profile for the GB system over the last 18 years, when demand has grown 18% then fallen 9%. The plant that is only required for one hour of the year (the 59th GW in 2012, or 51st GW in 1994) need only start once to cover the very highest peak demand. Similarly, all plants that can run for 8760 hours (the first 17–24 GW) do not have to start up (except for maintenance). Moving from these extremes to the mid-merit plants, which are required for 3,000–6,000 hours, the number of starts increases to approximately one per day.

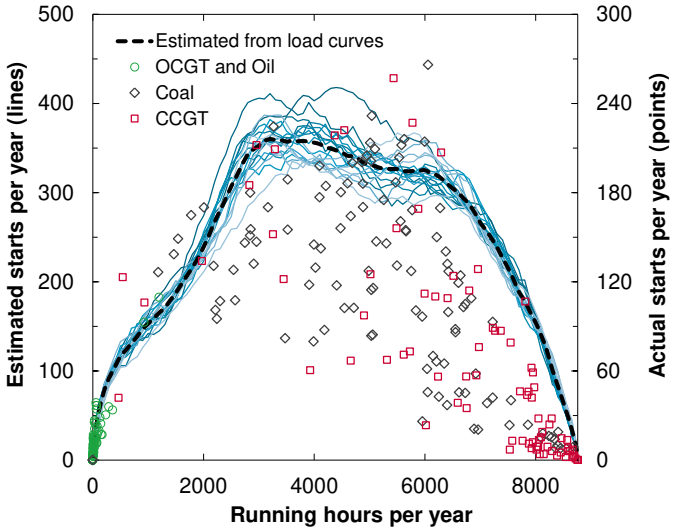

Fig. 4: The estimated number of plant start-ups for the GB system from 1994–2011. Thin lines show the estimations made from individual years' net load curves, and the period average highlighted as the thick broken line. The actual number of start-ups for individual plants during 2009-10 is shown as points (note the separate axis, which is scaled by a factor of 0.6).

This simple algorithm neglects the fact that it can be cheaper to reduce the output of several plants than to shut down a single plant, and so it over-estimates the number of start-ups. The actual behaviour of British stations during 2009–10 is shown in Figure 4 by the cloud of points. The number of running hours and start-ups was calculated from unit-level data collected from Elexon [29]. These were adjusted for units with long outages to estimate full-year values. For example, if a unit suffered a three month outage, its running hours and number of starts were multiplied by 1/0.75.

Scaling the predicted number of starts by 0.6 aligns the estimations from historic load curves (the lines in Figure 4) with the frontier of actual plant starts (the uppermost points). Some units sit below this frontier, starting up fewer times than is predicted by the rescaled algorithm. This is due to the diversity of operators' behaviour: some plants may preferentially be part-loaded so that they contribute to spinning reserve, or to reduce the mechanical wear incurred by thermal cycling. It could be argued that the efficiency penalty from part loading will still incur a cost of similar magnitude to shutting and restarting.

Our simulations show that this method provides a better approximation than ignoring starts completely. A more sophisticated algorithm that accounts for operator's decision making processes (such as Batlle and Rodilla's algorithm [24]) could be used in place of our simple load curve analysis to gain better results, albeit with higher data and computational requirements.

This curve of starts versus running hours can be reduced to an analytic form as in (1):

$$NS_h = n\, f \sin\left(\frac{h\,\pi}{8760}\right)^{\alpha} \qquad (1)$$

The number of starts (NS) for a given number of running hours (h = [1,8760]) equals the peak number of starts (n) found from the historic load data, multiplied by a scale factor (f), which reduces this peak to realistic levels. This is multiplied by the sine of the fraction of the year spent running, raised to a dimensionless power (α) which controls the duty cycle of the curve. The parameters n and α are

determined by performing a non-linear least-squares fit of the profile of start-ups (Figure 4) to a half-period of a sine wave.

Another approximation to the profile of start-ups is a harmonic series of sine waves. Between 7 and 11 harmonics are required for different countries to give a better fit than the two-parameter sine-power law proposed in Equation 2, and so this is adopted for simplicity. A further option is to use a lookup function to capture all the detail of Figure 4, albeit at the expense of model runtime.

By considering the number of starts against the number of hours per year the plant must run, we enable this method to be translated easily to systems other than the UK. Other systems, such as those in the US, France or Germany also show a consistent pattern within a country over time, but there are marked differences between countries due to the different extents of electric heating and pumped hydro storage employed. Based on gross load curves, before hydro output has been allocated to individual hours, stations in hydro-rich regions undergo fewer start-ups (lower n), with a flatter profile against operating hours (lower α). Table I gives representative values for the parameters of Eq. 1 derived from load curve data from several countries [30]. Scale factors are derived for the British and US systems from analysis of individual power station outputs [29, 31]; these cannot be calculated for the other countries as plant-level data are not available.

System	n	α	Years	f	Years
Great Britain	357	0.689	94–13	0.60	09–10
ERCOT	301	0.685	04–13	0.68	08–12
New England	376	0.687	80–13	0.53	08–12
PJM-E	359	0.804	99–13	0.62	08–12
Japan (TEPCO)	384	0.729	08–13	*n/a*	
Germany	343	0.578	07–13	*n/a*	
France	261	0.521	96–13	*n/a*	
Norway	183	0.472	07–13	*n/a*	

Table I: Start-up parameters estimated from load data.

3.2.1 Impact of Start-ups on Generating Cost

The total annual cost of start-ups can be calculated for each station type at each loading point by multiplying the number of starts per year by the cost per start (given later in Table II). The modified screening curves can be represented analytically as in (2):

$$TC_{g,h} = FC_g + VC_g\, h + c_g\, NS_h \qquad (2)$$

As with traditional screening curves: the total cost (TC) in £/kW-year for generator type g with h running hours per year equals the fixed costs (FC) plus variable costs (VC) times the number of hours run (h). Our modification adds the normalized cost of starting the generator (c_g) in £/kW given later in Table II, multiplied by the expected number of starts for that loading point (NS_h).

The total cost is used to calculate annual profits for each station type. As plant starts raise the cost of generation differently across the generator types and loading points, they alter the optimal mix of capacity.

3.2.2 Impact of Start-ups on Electricity Prices

The adjustment we make to hourly electricity prices differs from this in two ways. Firstly, the annual cost of starting the marginal generator type in each hour is divided by the number of running hours per year (h) to give the marginal cost of start-ups in £/MWh produced by that plant type. Another way to think of this is that we divide the cost per start by the average number of hours the plant is expected to run before stopping (h / NSh). This is the similar to the algorithms used to set prices in the British Electricity Pool in the 1990s and in the current Single Electricity Market in Ireland [32, 33].
Secondly, start-ups do not solely act to increase electricity prices. At local demand minima, prices are depressed if an increase in demand could avoid a plant having to shut down, and therefore start-up again. For example, the British Pool had separate pricing rules for so-called 'Table B' periods when demand was locally low. The increase in price is therefore calculated relative to the cost for the median running time experienced by all plants, as in (3).

$$PS_h = c_g\left(\frac{NS_h}{h} - \widetilde{\left(\frac{NS}{h}\right)}\right) \quad (3)$$

The price from start-ups for a given number of running hours (PS_h) that is added to electricity prices (in £/MWh) is equal to the cost of starting the marginal type of generator (c_g) normalized by its capacity (£/MW), multiplied by the difference between the number of starts per hour for the given loading point and the median number of starts per hour across all loading points.

This adjustment to prices assumes that generators are able to roll start-up costs into their energy bids, raising them above marginal cost. While this is at odds with a pure locational marginal pricing (LMP) system, we find that this heuristic works well for the current British market. It may not do so in other contexts.

4 Unit Commitment Capacity Optimiser (UCCO)

We compare our heuristic approach to the results from a generic, mixed-integer optimized dispatch model written in the GAMS language and controlled by a web interface. This model consists of standard equations, and is similar to those demonstrated in [2] and [3]. The model is formulated in the GAMS language, and is available to download from: http://hdl.handle.net/10044/1/12715.

The model consists of several classes of power stations that are dispatched to meet a set of time-varying demands so as to minimize the cost of generation. Demand rationing is achieved by inserting several tranches of demand that can be shed in response to high prices, which can be interpreted as a stepped demand curve, in contrast to the linear demand curve used in [7, 25]. In this setting, minimizing cost is equivalent to maximizing welfare, and it is a standard result in economics that this should be equivalent to the outcome that a perfectly competitive market would produce.

The model can optimize a fleet of 20 plant types over 8,760 hours of demand in around 20 minutes on a standard workstation (3 GHz, 4 GB RAM). Monte Carlo trials and multi-dimensional sensitivity studies can be conducted using the web interface, but can require a significant amount of computer time (e.g., 1,000 trials take a fortnight to solve).

Once plant characteristics, demand data and economic information are supplied, the model optimizes the scheduling and output of plants subject to the following constraints:

- Demand, plus a reserve margin, must be served by the operating plants; failure to do so incurs penalties due to curtailing supply (spilling wind) or demand, charged at the value of lost load (VOLL);
- Price sensitive demand is modelled with several tranches of consumers (usually large industrial users) that are able to reduce load in return for a scarcity price (between generators' marginal costs and VOLL);
- Plants have a minimum stable output below which they must shut down. Restarting the plant incurs a cost and time penalty;
- Plants have reduced efficiency when operating part loaded.
- Hydro and pumped storage are scheduled within the economic dispatch, subject to availability constraints due to water levels.

The model finds the short-run equilibrium – how to best operate a given set of plants so as to minimize the total cost of generation. The long-run equilibrium – the capacity of plants that would be best to build – is found when the profits of each type of plant are closest to zero, and so there is no incentive for new capacity to open or for existing capacity to retire.

The model is first solved as an integer problem to calculate the primary solution (giving plant schedules and outputs). The marginal cost of electricity is then found by fixing the number of units online and then relaxing the integer constraints on other variables to solve as a linear program.

GAMS is not capable of performing nested optimizations (a model within a model), so the web interface is used to iteratively search for the long-run equilibrium: testing a set of plants, refining levels of capacity based on their profits and re-testing until convergence is achieved. The plant with profits furthest from zero (after depreciation, a return on capital and all variable costs, including starts) has its capacity adjusted: upwards if it is making excess profit, downwards if it is making a loss. The process stops when all plants are receiving exactly the right revenues to cover their costs. Non-convexities in the generator cost function mean that a single long-run equilibrium may not exist. We find this method always reaches an equilibrium, but do not guarantee it is the global rather than a local optimum.

5 Scenarios for Validation

We use two scenarios based on the British electricity system to validate the hydro scheduling and start-up heuristics. First we run a historic scenario from 2009–10 to compare price formation against historic outturn. We then consider the system in 2020, when dynamic plant constraints have greater impact due to the increased capacity of intermittent wind. In this case, the optimal mix of thermal capacity chosen by the stack model is validated against that from the more complex unit commitment model.

Modelling these scenarios requires five sets of data: the capacities, technical limitations and costs of each station type, plus time series of demand and output from wind farms.

5.1 Historic Scenario: GB in 2009–10

Half-hourly demand data was taken from National Grid [30]. From this we subtracted the output from must-run CHP, interconnectors and wind [29], to give the net demand that dispatchable generators had to meet. For our first test, we also subtracted the output from hydro stations (1 GW of flow hydro and 2.4 GW of pumped storage) to give the actual demand that thermal stations met. Hydro output was added back to demand (and pumping consumption subtracted) for our second test, so that errors introduced by hydro allocation and the start-up heuristic could be separated from one another. Plant capacity and availability was fixed to historic levels based on Elexon data [29]. Nine types of station were modelled, as shown in Table II.

	Installed Capacity in 2010 (GW)	Annualised Capital Cost (£/kW-year)	No-load Cost (£/h)	Cold Start Time (hours)	Start-up Cost (£/MW)	Net Efficiency (LHV)	Total Fixed Cost (£/kW-year)	Total Variable Cost (£/MWh)
Nuclear	8.9	401	320	96	4,000	35.1%	470	8
Coal (3 subgroups)	26.0	207	620	4	200	33.5% 36.5% 39.5%	240	47 51 56
CCGT (3 subgroups)	25.3	85	4,900	2	50	50.5% 53.0% 55.5%	103	54 57 59
OCGT	0.7	58	130	0.1	10	32.3%	72	99
Oil	2.6	146	360	0.1	10	33.8%	188	141

Table II: Technical and economic parameters for UK power stations.

Coal and CCGT stations were split into three tranches in a fixed 25:50:25 ratio of capacity, with the range of efficiencies shown. These efficiencies were taken from environmental performance reports from the major operators and DUKES [34], and represented the mean ± one standard deviation.

Thermal plants operate with an availability of 80–90% (summer/winter) based on historic performance. Nuclear availability was 5% lower than this all year round. Wind output is modelled explicitly using a profile of resource availability, while hydro output was constrained by water availability, giving annual load factors of 42% and 15% for run-of-river and pumped hydro respectively.

Historic fuel and carbon prices for each quarter were taken from government statistics [35], and were implemented using eight separate stacks solving the two-year problem. We attempted to simulate historic output and prices. These were validated against actual wholesale prices for 2009–10 (the Market Index Price) obtained from Elexon [29].

5.2 Future Scenario: GB in the 2020s

Our demand time-series for 2020 was based on several years of National Grid data, from 1994–2011 [30]. We scaled each year's annual demand to 318 TWh, which is the level projected for the 2020s [26]. Historic demands were adjusted by the ratio of 318 TWh to their annual weather-corrected demand (as opposed to actual demand), so that we preserve hour-to-hour and year-to-year variation due to weather while removing fluctuations due to the level of economic activity.

The simulated wind output was subtracted from this demand, giving 18 years of load net of wind. In the stack models, these were combined into a single 8,760-hour load curve so that a range of weather-years were accounted for. In the UCCO model, all 18 years of data (157,752 hours) were used in full, with each year solved in parallel.

The optimal capacity mix depends on construction as well as operating costs. We derived plant costs from three studies specific to the UK and two internationally [26], aggregating their 2020 projections of annualized capital investment cost (defined as the annual rent required to cover overnight capital cost plus interest over the lifetime of the plant), fixed and variable operating costs, and thermal efficiencies. Fuel costs were based on the UK government's central scenario for 2020 [36]: £7.94 for coal, £37.07 for oil, and £23.37 for gas (per MWh of fuel). Carbon emissions are priced at £30 per tonne, which is the floor price established for 2020 under the UK carbon price support scheme [37].

No-load costs are derived from the intercept of total fuel cost against plant output, and represent the penalty of decreasing part-load efficiency. Based on data from US generators [31, 38], we assume that plant efficiency scales linearly with output, falling 6% from full to minimum output (for coal, OCGT and oil) or 16% (for CCGT and nuclear).

Start-up costs are derived from the cost of fuel required to heat the generator to temperature plus the cost of the carbon emitted. The wear and tear caused by start-ups is not factored in; however, this could increase the start-up cost significantly [39]. Shut-downs are considered to incur zero cost.
For this set of simulations we searched for the so-called "greenfield" solution; the long-run equilibrium capacity mix, assuming that there was no existing plant.

5.3 Wind Resource Data

The hourly output from the British fleet of wind farms was simulated using the Virtual Wind Farm model described in [40] and [41]. This takes hourly wind speed data from NASA, interpolates them to the location and height of the turbines at each individual wind farm, and then converts to power outputs using the power curve for the model of turbine installed at that farm (which varies between manufacturers and specific design features).

Figure 5 shows how this process validates against historic metered output data from the GB wind fleet, taken from Elexon [29]. The R^2 between simulation and reported output is over 0.95, and the root mean square error is 233 MW, implying that half-hourly output can be simulated with an accuracy of ±4.5%.

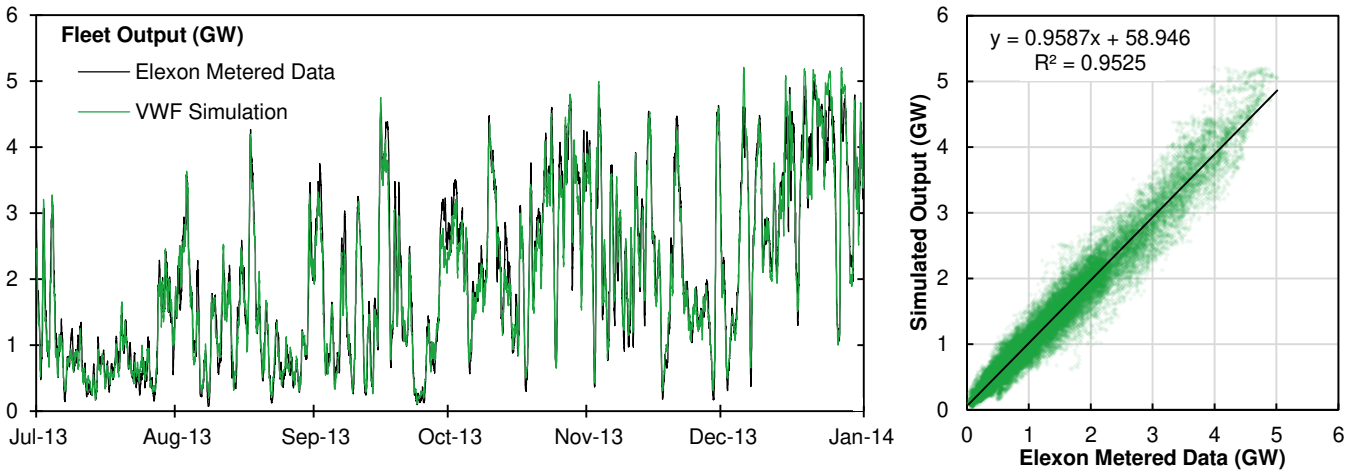

Fig. 5: Validation of the simulated wind farm output patterns from the Virtual Wind Farm model, showing a six-month sample from the transmission-connected GB wind fleet (46 of the largest farms) and the correlation over the two-year period 2012–13.

This model was used to simulate the output patterns from the fleet of wind farms that could be expected to be operating in 2020. We considered farms that are either operating, under construction or have obtained planning permission, giving 730 farms totalling 26 GW of capacity, to match National Grid's 'Gone Green' scenario [42] and projections by RenewableUK [43]. This gave a different pattern of demand from scaling up historic outputs due to the move towards more offshore farms in the North Sea. Our simulation was based on weather data from 1994–2011. This was the same period as our demand data, to preserve the correlation between weather and both demand and wind output.

The resulting wind fleet had an average capacity factor of 24.4% onshore, in line with historic averages [34, 44], and 34.7% offshore, which is two percentage points higher than historic experience, due to the anticipated move to larger turbines further from shore.

6 Results

6.1 Historic Comparison: GB in 2010

In the first stage of validation our stack model was calibrated to the GB system in 2009–10 and its results compared to historic output and price data. The model was configured to run with and without start costs, and with hydro outputs from actual outturn, the allocation from 30 tranches and from one tranche; so that the impact of each heuristic could be analysed separately.

6.1.1 Impact on Energy Output

With careful calibration of historic fuel prices and efficiencies, the stack model is able to predict the annual output of each station type very closely. The simulated capacity factors during 2009/10 were all within ±0.9% of historic levels: nuclear (78%), CCGT (70%), coal (44%) and OCGT (0.8%).

As explained in Section 3.2, the start-up heuristic does not affect dispatch and so had no impact on station outputs. The different methods of allocating hydro also had relatively little effect, altering capacity factors by only 0.3% at most. This is due to GB having relatively little hydro capacity (3.5 GW

in a 70 GW system), and the unbiased nature of the errors introduced by the allocation methods. Our proposed method of allocating multiple blocks gave an RMS error of ±301 MW from actual output, which is a 1.8-fold improvement over the simpler method of allocating one block (±550 MW).

When considering the hourly output by station type, the correlation between simulated and historic data yields R^2 values of 0.96 for CCGT and 0.93 for coal. OCGT output shows no correlation, as the exact hours in which peaking plants are used are poorly correlated. A model which emphasizes speed and simplicity cannot be appropriate for all applications, and this is apparently one of them.

6.2 Impact on Electricity Prices

We compare our simulated prices to the so-called Market Index Price (MIP), an average of the day-ahead and on-the-day prices which is less volatile than a real-time, or purely balancing price. The simple stack under-estimates annual average wholesale electricity prices by around 6%. The start heuristic adds £1.94/MWh, or 5.2% to the time-weighted wholesale price, bringing prices to within 0.9% of historic averages: £39.14/MWh (time-weighted) and £41.20/MWh (energy weighted).

Figure 6 shows the relationship between demand and price from the models. The simple stack shows the traditional stepped curve, with long horizontal segments when CCGT, coal and OCGT are marginal. The start heuristic introduces a slope to this curve because it allocates the cost of a start-up equally across all the hours that a station is generating. As demand increases, there are fewer hours to spread the cost of these starts over, and so prices gradually rise. This better replicates the historic price-demand relationship, shown as the solid black line in Figure 6.

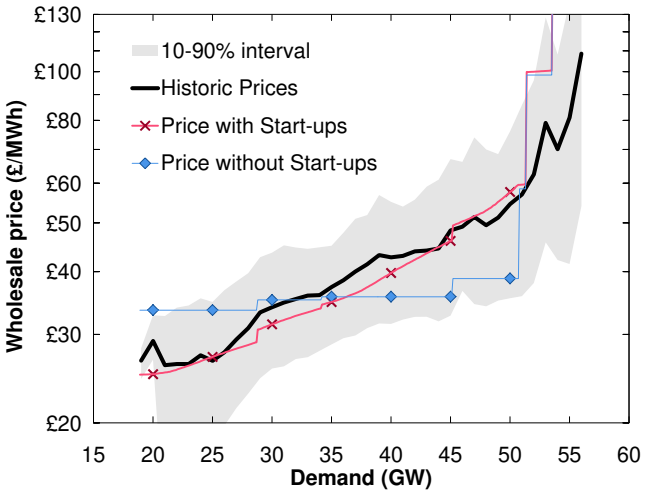

Fig. 6: Wholesale electricity prices produced by the simple and modified merit order stacks for the GB system in 2010, compared to historic.

The start heuristic depresses summer minimum prices by up to £8/MWh and adds up to £40/MWh to peak prices. For 90% of the year the change lies in the range of –£6 to +£16/MWh. This improves the correlation between simulated and historic prices, increasing the R^2 from 0.24 without to 0.39 with starts.

The R^2 value is improved but is still rather low, which may be expected from a simple deterministic model fed with high-level non-confidential data. A large spread is seen in historic prices: e.g., the

price at 30 GW can be the same as at 55 GW. In reality, fuel prices and efficiencies are not uniform across all stations, different costs are incurred for cold and hot starts, and unplanned outages can significantly influence prices.

6.3 Future Comparison: GB in 2020

In our second stage of validation the stack and unit commitment (UCCO) models were used to find the optimal thermal capacity mix for Britain in the 2020s. Two variants of the stack model were considered: the simple stack with no additional heuristics, and the modified stack with start-up costs and a 90% must-run constraint for the nuclear fleet.

6.3.1 Impact on Optimal Capacity Mix

The means by which start costs alter the optimal grid mix is demonstrated in Figure 7. The cost line for coal plants rises substantially in the mid-merit region as they cost 6 times more than other stations to start per kW of capacity. Similarly, the change to the OCGT cost curve is imperceptible due to its negligible start cost. [24] presents similar curves, plotted against loading point rather than operating hours. After correcting for this our curves are similar, except for there being no reduction in cost at high loading hours. The cost saving that baseload plants achieve by part-loading to avoid shutdown is factored into the screening curves in [24], whereas we make the simplification of accounting for it by post-processing prices. Nuclear costs show a different pattern, rising as utilization falls due to the cost of constraining wind in order to avoid extremely expensive shutdowns.

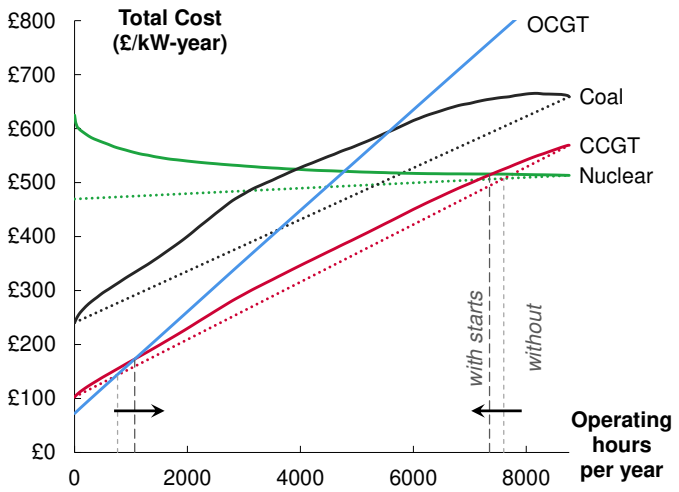

Fig. 7: Screening curves for four types of power station, showing the impact of start-ups on total annual cost. Dotted lines show the original cost curves, solid lines show the added impact of start-up costs. Vertical dashed lines indicate how the transitions between marginal stations (and thus the optimal running hours for each station type) shift when start-up costs are included.

The effect of these station-specific alterations is to move the transition points that give the optimal running hours for each type, and thus the optimal capacity. The vertical lines in Figure 7 show the transitions from OCGT to CCGT to nuclear being the cheapest station, which are moved inwards as the increase in cost for CCGT is larger than for the other two stations. This effect would be stronger if coal featured as one of the cheapest generators, but the assumed fuel and carbon prices prevent this.

Figure 8 shows the long-run equilibrium capacity mix chosen by each variant of the models. The mix chosen by the UCCO is taken to be the 'correct' answer as it fully accounts for dynamic plant constraints. The stacks produce similar results to the UCCO, deviating by at most 20%.

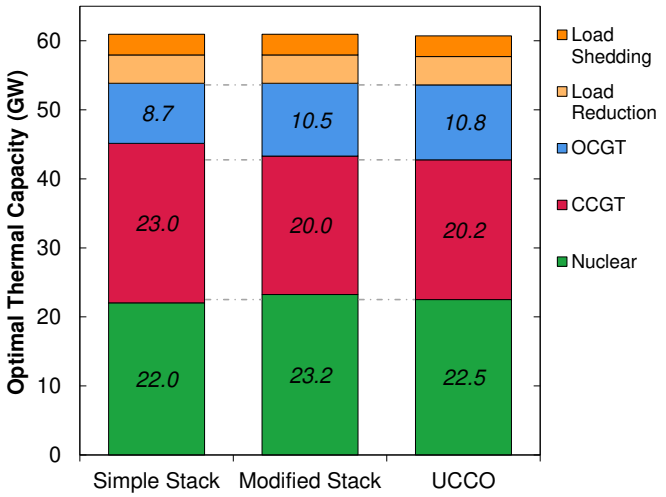

Fig. 8: Optimal capacity mixes calculated by each model.

The simple stack deviates furthest, over-predicting CCGT capacity by 2.85 GW and under-predicting OCGT by 2.15 GW. Adding a must-run constraint to nuclear plant had the expected effect of reducing the optimal capacity, as the additional cost of spilling wind (which was required when nuclear output could not be reduced any further) meant less nuclear capacity could compete with CCGT. Incorporating the cost of start-ups reversed this effect, reducing CCGT competitiveness against both nuclear and OCGT.

The modified stack therefore saw CCGT capacity squeezed at both ends because of its higher ratio of start-up cost to variable generation costs. The deviation from the UCCO results was therefore reduced from 2.85 to 0.20 GW for CCGT and from 2.15 to 0.30 GW for OCGT. The error in nuclear capacity was slightly increased, from 0.45 to 0.75 GW, which could be reduced by altering the must-run proportion of nuclear capacity.

The incorporation of start costs and must-run constraints reduced the RMS error in thermal capacity choices from ±2.1 to ± 0.5 GW in absolute terms, and from ±14% to ±3% relative to the installed capacity.

6.3.2 Impact on Energy Output

The constraints and modelling methods had more noticeable impacts on the levels of plant output, as plant run times were affected by their position within the stack (in addition to the changes in capacity). Table III gives the energy outputs predicted by the three models, showing that our corrections improve the accuracy of the stack model approximately two-fold.

The greatest deviations from the fully optimized result were at the extremities of the stack: the level of wind spilling at the bottom and the levels of OCGT output at the top. In all cases, the amount of load shedding was very similar by design, as it was determined primarily by the total amount of physical capacity installed.

	UCCO	Modified Stack		Simple Stack	
Nuclear	192.7 TWh	198.6	*(+3%)*	177.5	*(–8%)*
CCGT	86.4 TWh	81.4	*(–6%)*	102.6	*(+19%)*
OCGT	5.0 TWh	4.4	*(–11%)*	3.6	*(–28%)*
Wind spilling	–1.8 TWh	–2.0	*(+15%)*	–1.2	*(–30%)*
Load shedding	–80 GWh	–79	*(–1%)*	–79	*(–1%)*

Table III: Comparison of plant outputs for different modelling methods, highlighting the deviation from the fully optimized results.

6.3.3 Impact on Electricity Prices

The modifications to the stack have a similar impact on the distribution of electricity prices as in the historic comparison (Figure 6). Prices are raised during periods of high demand and lowered during baseload periods over 4,000 running hours. Figure 9 shows the price-duration curves resulting from the three model runs.

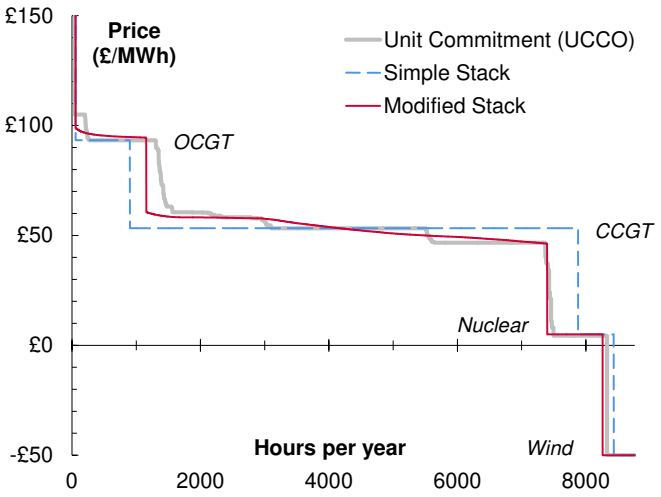

Fig. 9: Price-duration curves from the three models for the 2020 scenario.

The vertical segments of these curves indicate transitions between different generators being marginal (as labelled in the figure). Differences in the position of these transitions can be seen between the UCCO and the simple stack model due to the stack's error in both the capacity installed and its runtime. These errors are reduced significantly when adding the modification for start-ups.

The simple stack produces price-duration curves with long horizontal segments, as price is determined solely by incremental fuel costs. The start-up modification imposes a downwards slope on these segments, for the reasons given in Section 6.1.2. Within the range of 2,000–7,000 hours when CCGT is marginal, this heuristic appears to match the profile of start-costs and avoided shut-down savings predicted by the UCCO model at least in magnitude, although the finer structure (with several grades of cost adjustment) is simplified.

7 Conclusions

We propose and validate a new heuristic for incorporating start-up costs in the merit order stack, focusing on its ability to model electricity prices and the long-run equilibrium capacity mix. This heuristic is designed for speed and simplicity: it has no adverse effect on computational efficiency, and requires as a minimum only the net load curve (demand minus renewable output), and data on plant start costs.

By increasing peak electricity prices and reducing them during times of low demand, this heuristic improves the correlation with historic prices that can be obtained by a merit order stack by 60%. By inflating the cost of mid-merit plants, start-up costs move the crossover points at which OCGT or nuclear plant become the lowest cost generator, increasing the shares of these plant in the long run equilibrium mix. These changes are replicated in the price duration curve, enabling the modified stack model to approach the results of the fully optimized dispatch model.

We also demonstrate a marginal improvement on the peak-shaving algorithm for allocating hydro output. By splitting hydro capacity into multiple tranches with different utilizations, the historic distribution of output is better represented, but this is seen to have limited impact on results for the British electricity system due to the low levels of hydro installed.

These heuristics are demonstrated within a simple cost-minimizing model that assumes a perfectly competitive market. It would be possible to apply them to other situations where market power is exerted, using the start-up heuristic to generate improved cost data for supply function, Cournot or bi-level models [45].

Whether the results from this simple technique are acceptably accurate depends on the specific application and user's preferences. However, even when considering scenarios with a challenging level of intermittent renewable generation, it seems that there is still merit in an improved merit order stack.

8 Acknowledgements

We thank W. Wiesemann, Prof. A.J. Conejo and our anonymous referees for their thoughtful insights that have improved our work.

9 References

[1] N. Troy, E. Denny and M. O'Malley, 2010. *Base-Load Cycling on a System With Significant Wind Penetration*. IEEE Transactions on Power Systems, **25**(2): pp.1088-1097

[2] G. Strbac, M. Aunedi, D. Pudjianto, P. Djapic, F. Teng, A. Sturt *et al.*, 2012. *Strategic Assessment of the Role and Value of Energy Storage Systems in the UK Low Carbon Energy Future*. London, UK: Carbon Trust. http://tinyurl.com/c7gubyk

[3] A. Mills and R. Wiser, 2012. *Changes in the Economic Value of Variable Generation at High Penetration Levels: A Pilot Case Study of California (Report LBNL-5445E)*. Berkeley, CA: Lawrence Berkeley National Laboratory.

[4] S.E. Stoft, 2002. *Power System Economics: Designing Markets for Electricity*. Chichester, England: Wiley.

[5] D. Kirschen and G. Strbac, 2004. *Fundamentals of Power System Economics*. Chichester, England: Wiley.

[6] S. Borenstein, J.B. Bushnell and F.A. Wolak, 2002. *Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market.* American Economic Review. **92**(5): pp. 1376-1405.

[7] S. Borenstein, 2005. *The Long-Run Efficiency of Real-Time Electricity Pricing.* Energy Journal. **26**(3): pp. 93-116.

[8] A.D. Lamont, 2008. *Assessing the Long-Term System Value of Intermittent Electric Generation Technologies.* Energy Economics. **30**(3): pp. 1208-1231.

[9] J. Bushnell, 2010. *Building Blocks: Investment in Renewable and Nonrenewable Technologies*, in *Harnessing Renewable Energy*, B. Moselle, J. Padilla, and R. Schmalensee, Editors. Washington, DC: RFF Press.

[10] C. De Jonghe, B.F. Hobbs and R. Belmans, 2012. *Optimal Generation Mix With Short-Term Demand Response and Wind Penetration.* IEEE Transactions on Power Systems, **27**(2): pp. 830-839

[11] F. D. Munoz, B. F. Hobbs and S. Ka, 2012. *Efficient Proactive Transmission Planning to Accommodate Renewables*, Power and Energy Society General Meeting, San Diego, CA.

[12] F.D. Munoz, B.F. Hobbs, J.L. Ho and S. Kasina, 2014. *An Engineering-Economic Approach to Transmission Planning Under Market and Regulatory Uncertainties: WECC Case Study*. IEEE Transactions on Power Systems, **29**(1): pp. 307-317

[13] A. Sturt and G. Strbac, 2012. *Efficient Stochastic Scheduling for Simulation of Wind-Integrated Power Systems*. *IEEE Transactions on Power Systems*, **27**(1): pp. 323-334

[14] P. Meibom, R. Barth, B. Hasche, H. Brand, C. Weber and M. O'Malley, 2011. *Stochastic Optimization Model to Study the Operational Impacts of High Wind Penetrations in Ireland*. IEEE Transactions on Power Systems, **26**(3): pp. 1367-1379

[15] S. Jin, A. Botterud and S.M. Ryan, 2013. *Impact of Demand Response on Thermal Generation Investment With High Wind Penetration*, IEEE Transactions on Smart Grid, **4**(4): p. 2374-2383

[16] R. Green, I. Staffell, and N. Vasilakos, 2014. *Divide and Conquer? k-means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System.* IEEE Transactions on Engineering Management, **61**(2): pp. 251-260.

[17] S. Wogrin, J. Barquín and E. Centeno, 2013. *Capacity Expansion Equilibria in Liberalized Electricity Markets: An EPEC Approach*, IEEE Transactions in Power Systems, **28**(2): pp. 1531-1539

[18] S. Wogrin, J. Barquín and E. Centeno, 2013. *Generation Capacity Expansion Analysis: Open Loop Approximation of Closed Loop Equilibria*. IEEE Transactions in Power Systems, **28**(3): pp. 3362-3371

[19] N. Langrene, W. van Ackooij and F. Breant, 2011. Dynamic Constraints for Aggregated Units: Formulation and Application. IEEE Transactions on Power Systems, **26**(3): pp. 1349-1356

[20] B.S. Palmintier and M.D. Webster, 2014. *Heterogeneous Unit Clustering for Efficient Operational Flexibility Modelling*. IEEE Transactions on Power Systems, **29**(3): pp. 1089-1098

[21] C. De Jonghe, E. Delarue, R. Belmans and W. D'haeseleer, 2011. *Determining Optimal Electricity Technology Mix with High Level of Wind Power Penetration*. Applied Energy, **88**(6): pp. 2231-2238

[22] S. Jin, A. Botterud and S.M. Ryan, 2014. *Temporal Versus Stochastic Granularity in Thermal Generation Capacity Planning With Wind Power.* IEEE Transactions on Power Systems, **29**(1): pp. 1-9

[23] A. Shortt, J. Kiviluoma, and M. O'Malley, 2013. *Accommodating Variability in Generation Planning*. IEEE Transactions on Power Systems, **28**(1): pp. 158-169

[24] C. Batlle and P. Rodilla, 2012. *An Enhanced Screening Curves Method for Considering Thermal Cycling Operation Costs in Generation Expansion Planning*. IEEE Transactions on Power Systems, **28**(4): pp. 3683-3691

[25] R.J. Green and N. Vasilakos, 2011. *The Long-Term Impact of Wind Power on Electricity Prices and Generating Capacity*. CCP Working Paper 11-4, Centre for Competition Policy, University of East Anglia. http://ssrn.com/abstract=1851311

[26] R. Green and I. Staffell, 2013. *The Impact of Government Interventions on Investment in the GB Electricity Market*. European Commission State Aid Decision SA.34947 on UK Electricity Market Reform – Contract for Difference for the Hinkley Point C New Nuclear Power Station.

[27] S. Borenstein and J. Bushnell, 1999. *An Empirical Analysis of the Potential for Market Power in California's Electricity Market*. Journal of Industrial Economics, **47**(3): pp. 285-323.

[28] R. Billinton and R.N. Allan, 1984. *Reliability Evaluation of Power Systems*. London: Pitman Books.

[29] Elexon, 2014. *Balancing Mechanism Reporting System: FPN, FUELHH, UOU2T14D and MIP tables*. http://bmreports.com/

[30] Data from various ISO and TSO websites. http://tinyurl.com/l3t7c7r

[31] EPA, 2014. *Air Markets Program Data*. http://ampd.epa.gov/ampd/

[32] Electricity Pool, 1999. *The Pool Rules*. London, Electricity Pool of England and Wales.

[33] SEMO, 2014. Pricing & Scheduling Frequently Asked Questions. http://tinyurl.com/pr4r7dv

[34] I. MacLeay, K. Harris and A. Annut, 2013. *Digest of United Kingdom Energy Statistics (DUKES)*. http://tinyurl.com/decc-dukes

[35] Department of Energy & Climate Change, 2014. *Quarterly Energy Prices*. http://tinyurl.com/decc-qep

[36] Department of Energy & Climate Change, 2014. *Fossil Fuel Price Projections*. http://tinyurl.com/decc-ffpp

[37] HM Treasury, 2011. *Carbon Price Floor Consultation: The Government Response*. http://tinyurl.com/96dwppr

[38] J.B. Bushnell and C. Wolfram, 2005. *Ownership Change, Incentives and Plant Efficiency: The Divestiture of U.S. Electric Generation Plants*. University of California Energy Institute.

[39] O. Rosnes, 2008. *The Impact of Climate Policies on the Operation of a Thermal Power Plant*. Energy Journal, **29**(2): pp. 1-22.

[40] I. Staffell and R. Green, 2014. *How Does Wind Farm Performance Decline with Age?* Renewable Energy, **66**: pp. 775-786.

[41] R. Green and I. Staffell, 2014. *How Large Should a Portfolio of Wind Farms Be?* USAEE Working Paper No. 14-182. http://ssrn.com/abstract=2529791

[42] National Grid, 2013. *UK Future Energy Scenarios*.

[43] RenewableUK, 2014. *UK Wind Energy Database (UKWED)*. http://tinyurl.com/reuk-ukwed

[44] Ofgem, 2014. *Renewables and CHP Register*. http://tinyurl.com/ofgem-rocs

[45] S.A. Gabriel, A.J. Conejo, J.D. Fuller, B.F. Hobbs and C. Ruiz, 2013. *Complementarity Modelling in Energy Markets*. New York: Springer.