Are the British Electricity Trading and Transmission Arrangements Future-proof?

Richard Green

Institute for Energy Research and Policy,
University of Birmingham
Birmingham B15 2TT, UK
Tel: +44 121 415 8216
Fax: +44 121 414 7377
r.j.green@bham.ac.uk

Abstract

In Great Britain, electricity is traded in an energy-only market that relies upon bilateral trading until shortly before real time. The GB System Operator also uses bilateral trading to respond to changes in demand and generation and resolve transmission constraints. Prices are not explicitly spatial, although well-placed generators can charge the system operator more for their output. This paper argues that these arrangements are not well-suited for the challenges of accommodating nearly thirty percent of intermittent wind generation, often located far from demand. The market design already implemented in the north-eastern United States is likely to be more efficient.

1. Introduction

The electricity industry in Great Britain was liberalised in 1990, with a compulsory day-ahead spot market covering England and Wales. This Electricity Pool was abolished in 2001, and replaced by the New Electricity Trading Arrangements, which evolved into the British Electricity Trading and Transmission Arrangements (BETTA) in 2005, when they were extended to cover Scotland.

The guiding principle of NETA's (and hence BETTA's) design was that electricity should be treated as much like a "normal" commodity as possible, while still recognising the physical characteristics of electricity. This means that there is a balancing mechanism run by the system operator, National Grid, to ensure that demand and generation are kept in balance and transmission constraints are respected, but no other market was centrally organised. Instead, most electricity was traded bilaterally (or internally, for integrated firms), with some trading on electronic exchanges to aid transparency. A voluntary day-ahead auction has recently been introduced, but its turnover is low.

The introduction of NETA was controversial, with many academic commentators suggesting that the design was inferior to that of the Pool it replaced (Hogan, 2002; Newbery, 1998; Wolfram, 1999). Despite this, the market has operated smoothly for more than eight years, with only one significant failure to match generation to demand.

The question for this paper, however, is whether BETTA is well-suited for the challenges of the next decade or so. The UK has accepted a target for renewable energy of 15% of its final energy consumption, as part of the European Union's strategy of providing 20% of its energy from renewable sources by 2020. Because it will be harder (or more expensive) to absorb this proportion of renewable energy in the (larger) sectors of heat and transport, it is likely that more than 30% of electricity will have to be from renewables if the UK is to meet its target. Most of this is likely to come from wind energy.

Wind farms (and some other renewable energy generators) are intermittent sources, dependent on the strength of the wind. When the amount of wind capacity is low, the variation in its output can just be seen as equivalent (if opposite) to variation in demand, and does not need to be separately managed. With a large amount of wind capacity, however, the variation in its output must be explicitly managed, determining the amount of spinning reserve that the system controllers need to keep available, in case a sudden decrease in the wind raises the amount of electricity needed from other stations. Furthermore, very little of the wind capacity can be relied on to meet the peak demand for power, and so the industry's total capacity will need to reflect this, with many conventional power stations that rarely run, but are needed in case high demand for power coincides with low wind speeds.

These are problems that can be managed, and while their costs are significant (Gross et al, 2006) they are not insurmountable. The question for this paper is whether the current trading

arrangements used in Great Britain will be fit for purpose in this changed environment. From this point of view, the trading and transmission arrangements must fulfil three key functions.

- First, they must give generators the incentive to build new plants (or keep old ones open) if the capacity is needed, and not if it is not.
- Second, they must make it possible for generators to connect plants to the grid in a timely manner, provided that the system can cope with the station's output.
- Third, they must promote efficient operation by the stations connected to the grid, so that the cheapest stations available generally produce as much as possible, while respecting transmission constraints and providing an adequate reserve for unexpected changes.

The next section of this paper describes the current market and transmission access rules in Great Britain. The third section discusses renewable generators and how they are supported in the UK. Section 4 discusses the challenges involved in creating and operating a low-carbon electricity system with a high proportion of renewable generation. The succeeding sections ask whether the current market rules are well-suited to meet the three requirements set out above, and conclude that they are not. The market design already used in parts of the United States offers a model which is more likely to achieve these tasks. A brief conclusion sums up.

This paper does not describe the US market design in any detail – descriptions are available in many papers, including Bowring (2006) and Hogan (2002). The focus of the paper is on the UK, but most EU countries use a similar system of power trading, in that the price of energy is the same across the whole country, and if transmission congestion occurs, the system operator has to resolve this by buying and selling power, outside the main markets. To the extent that other EU countries have (or will have) large amounts of intermittent renewable generation, and that it is concentrated in particular areas remote from the main centres of demand, then they will face similar problems to those that this paper predicts the UK will have to deal with.

2. The British Electricity Trading and Transmission Arrangements

The formal rules of BETTA are set out in codes which cover different aspects of the relationship between generators, retailers and the transmission companies. While most energy trading is bilateral, between generators and retailers, imbalances between their contractual and physical positions are inevitable, and the Balancing and Settlement Code sets out how these will be dealt with. It also contains the rules for the Balancing Mechanism, through which National Grid, as

system operator, buys and sells power to keep the system secure.[1] Changes to the codes are proposed by one or more companies, endorsed (or not) by an industry panel, and decided by the industry's regulator (the Gas and Electricity Markets Authority), with the possibility of appeal if the regulator goes against the panel's recommendation.

While most electricity trading is bilateral, "normal" bilateral trading must stop at Gate Closure, currently one hour before the start of each half-hour trading period. Companies notify the system operator of their intended physical and contractual positions, having already provided indicative information to help with operational planning. (To help with operational planning, they will already have provided indicative information on these.) From this point onwards, only the system operator can initiate trades.

The system operator must ensure that demand and generation are kept in balance, and transmission constraints are respected, and does so by trading through the Balancing Mechanism. Generators can volunteer offers to supply power to the market, and bids to buy from it (which involve generating less), although participation is not compulsory, and some generators will submit prices at levels which makes it very unlikely that they will be accepted. Large customers can also offer to supply power by reducing their demand on request.

The system operator will have to increase generation (and hence accept offers to sell power) if demand is unexpectedly high or if there are plant failures that reduce output. To prepare for these, the system operator may also arrange for some stations to run part-loaded, able to increase output very rapidly, by selling back some of their output – it will then need to buy a similar amount of power from other stations to balance supply and demand. If there is a transmission constraint, because the bilateral trades imply greater flows over one or more transmission lines than those lines can safely accept, then the system operator will also have to sell back some generation on the exporting side of the constraint, and buy an equal amount of power on the importing side.

At the end of each half-hour, the system operator's total sales and purchases are added up. If the system operator has had to buy more power than it sold back, then the system was short, and a System Buy Price will be calculated. If the system operator was a net seller, then the

[1] The other codes are the Grid Code, concentrating on technical requirements; the Connection and Use of System Code, which governs commercial terms for using the transmission system; and the System Operator – Transmission Owner Code, which governs the relationship between National Grid (in its role as GB System Operator) and the three companies which own transmission in Great Britain (National Grid itself, Scottish Power, and Scottish Hydro (part of Scottish and Southern Energy)).

system was long, and a System Sell Price will be computed. To compute the System Buy (Sell) Price, the most expensive purchases (cheapest sales) are "tagged out" against the same volume of sales (purchases) – this is so that trades made to keep the system operating smoothly (providing spinning reserve and resolving constraints) do not affect the imbalance prices. The System Buy (Sell) Price is then based on the average cost (revenue) of the next 500 MW of purchases (sales). Any market participant who is short of power when the system as a whole is short has to pay this price for its imbalance. If a company has a surplus of power when the system is short, however, this is likely to help the system, and they now receive an imbalance price is now based on recent trades reported in the short-term markets. In other words, if the industry turns out to be short of power, companies who need to buy more to cover an imbalance have to pay the price of the system operator's purchases made on their behalf in the balancing mechanism, but those who had power to dispose of receive the same price that they might have been able to obtain in the short-term markets. Buyers will always pay the System Buy Price, and sellers receive the System Sell Price, but the prices are calculated in different ways, depending on the overall balance.

National Grid does not rely purely on the balancing mechanism to keep the system running smoothly. It organises regular tenders for reserve plant and for other ancillary services. Some of the reserve is offered by large consumers – these tenders have been more successful in attracting demand side participation than the balancing mechanism.

It should be noted that all electricity trading is carried out at a (notional) National Balancing Point, and so there is no geographical element to the System Buy Price, the System Sell Price or to bilateral trades. There is an implicit geographical influence on trading in the Balancing Mechanism, in that the system operator will sometimes have to pay a relatively high price to a generator which is particularly well-placed to resolve a transmission constraint. A generator which needs to buy back its power because it is on the exporting side of a constraint may be able to submit a particularly low price, but this signal of the low value of electricity at this site actually allows the generator to make an additional profit, assuming that it had earlier received full payment for its output. This is the disadvantage of resolving congestion via a counter-trading mechanism.

Congestion can be minimised if new generators are only connected to the transmission system when it is able to accept their output under most operating conditions. This is the

principle that was followed by the system operator in offering connections in Great Britain. The approach has been nick-named "invest then connect", and requires applicants to wait until any reinforcement has been completed, so that adding the station to the grid would not lead to an increase in constraints. The problem with the way this methodology has been applied is that it is based on the idea that the grid should be able to accept power from all the stations connected to it simultaneously. It has not taken (sufficient) account of the fact that when the output from wind generators is high, this will crowd out production from conventional stations and the grid will not need to be able to accept both.

The problem of the so-called "GB queue" has received political attention, since renewable generators wishing to connect to the grid in Scotland had been offered connection dates as late as 2023 (DECC, 2009a). As a short-term measure, National Grid and the regulator agreed in 2008 to abandon their "first come-first served" approach, under which the earliest station to request capacity is given the first chance to connect to the grid once it is available. Instead, the company would prioritise those stations that appear best-positioned to actually use the capacity – in other words, those able to being construction first.

The industry was unable to agree longer-term reforms, with competing proposals following the principle of "connect and manage". This would allow a generator to connect to the grid as soon as any local reinforcement works were completed, even if this would lead to an increase in constraints further away. The industry failed to agree whether these costs should be socialised (charged to all consumers and generators) or targeted on generators in the area behind the constraint. Not unnaturally, generators who wish to be (or are) located behind those constraints favoured a socialised approach, whereas most incumbents favoured targeting. When the industry failed to make progress at a speed acceptable to its regulator, the matter was passed to the government for an imposed solution. The government chose to socialise congestion costs, on the basis that this gave the best returns to renewable generators, and meeting the government's target for renewable energy was more important than the possible increase in congestion costs that the socialised approach would entail.

3. Policy for renewable generation in Great Britain

The UK has a target of providing 15% of its energy from renewable sources in 2020. The government's low carbon transition plan (DECC, 2009b) envisages that 30% of the country's

electricity should come from renewable generators in 2020. The incentives to build the necessary capacity will not primarily come from the electricity market discussed in this paper – renewable generators receive substantial support from a feed-in tariff (for smaller schemes below 5 MW in capacity) and from the so-called Renewables Obligation. This is effectively a tradable green certificate scheme, which requires electricity retailers (suppliers, in British parlance) to surrender Renewables Obligation Certificates (ROCs) equal to a set proportion of the power they sell, or pay a buy-out charge. The buy-out payments are actually recycled to the retailers who have surrendered ROCs, in proportion to the number they surrender, which gives the certificates a value which increases as the amount of renewable generation falls short of the target (since each surrendered certificate attracts a greater amount of recycled buy-out payments), thus strengthening the incentive to invest. Renewable generators receive ROCs for each MWh of power they produce, but the more expensive technologies can receive more than one ROC per MWh – wave and tidal generators receive 2 ROCs per MWh generated. In contrast, generators using the gas given off by landfill (waste disposal) sites receive only 0.25 ROCs per MWh.

Renewable generators that receive ROCs must also sell their power in the wholesale market for whatever price they can get. The market rules are thus clearly relevant to the amount of revenue renewable generators can expect to receive, and hence the incentive to invest in renewable plant, but the government has the option of increasing the specific support given to renewable generators, rather than adjusting the general market rules, should the UK appear to be in danger of missing its targets. For example, the government might raise the number of ROCs to be surrendered per MWh of electricity sold, which (by increasing the amount of buy-out payments to be recycled) would raise their value. To support particular technologies, the government could also award them more ROCs per MWh of generation[2] - which would have to be matched by an increase in the number of ROCs to be surrendered unless this support was to come at the expense of other technologies.

The possibility of increasing the support given to renewable generators makes it possible to separate decisions on market design from the question of whether we will meet our renewable generation targets. The issues are whether changes to the market design will increase or decrease the resource cost of meeting those targets, and how they will affect customers' payments and

[2] In an example of this, a review of the increased costs of offshore wind turbines led the government to announce that offshore wind stations accredited to join the scheme between April 2010 and March 2014 will receive 2 ROCs per MWh of electricity generated throughout their twenty-year period of eligibility, instead of the 1.5 ROCs per MWh received by older stations (DECC, 2009c).

generators' rents. The broad thrust of this paper is that the price mechanism should be used to manage congestion, which is likely to reduce the net revenues of generators located in areas far from the market, such as Scotland. Some projects, such as those at particularly windy sites, will go ahead anyway, and the lower revenues simply mean lower economic rents. Other projects would be deterred by a change in market rules that lowered their (total) net revenues below their costs. In normal circumstances, this would imply a more efficient use of resources, assuming that the new (and insufficient) revenues were a true reflection of the value of the project's output. The problem is that if these projects are needed to meet the renewables targets, then the support given to those generators will have to be increased to offset the impact of the change in the market rules. Unless the change in support can be targeted on those generators that really need it, consumers (or taxpayers) may have to increase their total payments by many times the amounts received by the marginal projects, and the extra money will increase the rents received by infra-marginal generators. This should be seen, however, as an argument for targeting support, rather than for deliberately distorting the market rules to keep down the costs of untargeted support schemes.

4. Challenges from renewable generation

Regardless of the extent to which generators depend on the wholesale market for their revenues (and smaller generators receiving the feed-in tariff will be completely insulated from it), their output will affect it. This matters, because it is quite likely that the electricity system in Great Britain will have to absorb 30 GW of wind generation capacity if the UK is to meet its renewable energy target for 2020 (House of Lords, 2008, DECC, 2009b). The fluctuations in output that this level of wind power would produce will be significantly larger than the UK has experienced in the past. As a proportion of the local demand, they would be on the same scale as those produced by wind energy in Western Denmark, with the important difference that Denmark has a very large interconnection capacity to Norway and Sweden (with their largely hydro-electric power systems and ability to store water instead of generating) and to Germany.

These fluctuations will affect the industry in two main ways. First, the industry's conventional and nuclear plants will experience much greater hour-to-hour changes in demand, and will have to be operated to respond to these. Figure 1 gives two duration curves for

simulations of these changes for 2020, based on data from Green and Vasilakos (2009).[3] That paper took historic hourly demand and weather data for 13 years (1993 to 2005), scaled the underlying demand to possible 2020 levels, and estimated the output that 30 GW of wind generators distributed across Great Britain and its territorial waters, would produce. While the dataset contains 13 sets of 8760 observations (and a few leap days), the curves presented here are scaled to the 8760 hours in a normal year.[4] The appendix gives more detail on how these figures were derived.

Figure 1. Hour-to-hour changes in electricity demand in Great Britain.

The thinner line is based on the change in simulated gross demand between adjacent hours – that is, in the total load on the electricity system (assuming no demand response to the prices that might be produced). These are effectively the changes that the industry has been coping with in the past (and could presumably continue to cope with), adjusted for the rising level of demand (which should be matched by rising capacity, and hence ability to change output). The thicker line shows the hour-to-hour changes in the simulated net demand – that is the gross demand for electricity, less the simulated level of wind output. Figure 1 clearly shows that the distribution of hour-to-hour changes in this net demand has been stretched – the absolute change at any point in the distribution is now greater. This is particularly pronounced at the extremes – the greatest hourly changes have risen to – 13.7 GW and + 17.4 GW, compared to changes in the level of gross demand that never go beyond – 6.3 GW and + 10.1 GW.

What the figure does not show is the balance between anticipated and unanticipated changes in the net thermal demand. If the changes are anticipated, then the market needs to provide incentives for generators to run in a manner that matches the pattern of demand for their output, but does not need to deal with sudden changes to this pattern. To the extent that the changes in demand are not predictable, however, the market rules need to ensure that (enough)

[3] The focus of that paper was on the range of market prices that might be produced, considering the impact of intermittent wind generation and of market power. The paper used a supply function model to produce simulated prices, however, and abstracted from the issues of market design considered here.

[4] Effectively, once the calculations have been performed, the observations for all 13 years are ranked in order and every 13th observation is used in these charts.

generators will change their outputs quickly, in order to respond to the new pattern of demand. This could be a significantly greater challenge.

The scale of the second problem that comes from variable renewable output can be seen in figure 2. This shows the industry's load-duration curves, ranking hours in order of increasing demand. The upper curve shows the number of hours for which the gross demand exceeds a given level, whereas the lower curve shows the duration of demands net of wind output, the load on conventional and nuclear plants. Figure 3 gives a detail of the top end of figure 2.

Figure 2. Load-duration curve for Great Britain.

Figure 3. Load-duration curve for Great Britain (detail).

It is clearly the case that the number of hours for which net demand will be between 55 GW and 65 GW has fallen significantly, even though the very peak demand is only slightly lower after subtracting the wind output. The total capacity needed to meet the peak demand is therefore almost unchanged, but the amount of capacity that is only needed for a few hours a year, on average, is much higher. The problem for market design is that this capacity will need to be properly remunerated, even though it will only run for a few hours a year, on average (less in some years, and more in others).

The alternative would be to use some mechanism to ration demand to a lower level of available capacity. The price mechanism is generally the most efficient way of doing this, giving some consumers a financial incentive to reduce their demand that reflects the cost of meeting it. System operators can also reduce the level of operating reserve that they hold (at the cost of raising the risk of a serious failure), serve less demand by supplying power at a lower voltage and, in extremis, cut off some customers. This may be a more efficient outcome than building so much capacity that it can always meet the (non price-sensitive) demand, but it should be clear that applying the price mechanism to demand could be expected to produce better results.

The variability of wind output thus leads to two challenges – adjusting the output from conventional plant to match the more variable net demand for it, and paying for a greater volume of rarely needed reserve plant. Gross et al (2006) have estimated that if up to 20% of British electricity came from variable renewable sources, the first of these would cost £2-3/MWh of

renewable output, and the second would cost £3-5/MWh. The figures (even per MWh) are likely to increase with the proportion of renewable output – National Grid have suggested that short-term balancing would cost the equivalent of £3-7 per MWh if the 2020 target is met (House of Lords, 2008).

A third challenge comes, not from the variability of renewable output per se, but from its likely locations. Renewable resources are not evenly distributed around the UK. The best sites for onshore wind generation are in the north and the west, as are some offshore stations. The bulk of these, however, will be in the North Sea, where shallower depths compensate for lower wind speeds than in the Western Approaches and the North Atlantic.

This means that parts of the network will see some very large inflows of renewable power, and these are (mostly) a long way from the centres of demand in the Midlands and south of the UK. If these inflows were constant over time, it would clearly be economic to reinforce the transmission system to handle the increased flows, or (if this was not possible) to retire other local generation to ensure that the lines are not actually constrained. In practice, however, the flows vary over time, and the lines will not always be constrained. It may not be economic to reinforce a transmission line that is only constrained for part of the time, or to close a station that is often able to get its power to consumers. In that case, constraints on the system will increase, both in terms of number and the amount of energy that is affected.

Figure 4. Load-duration curve for Scotland.

Figure 4 gives an impression of this, subtracting the predicted wind output in Scotland from the simulated demand level for 2020. (As described in the appendix, this is based on four years for which we have detailed demand figures for Scotland and wind data, rather than the thirteen years underlying the other figures.) For around 1,250 hours a year, the wind generation exceeds the demand in Scotland, implying that some power will have to be exported across the interconnector with England. The capacity of that interconnector is due to rise to 3 GW, which implies that there will be 140 hours a year (on average) in which it would not be possible to export all the available power to England. However, we should also consider the output from the 2 GW of nuclear stations in Scotland – if they are still running on base load in 2020, this would

mean that exports of power were constrained as soon as the output from wind power exceeded local demand by 1 GW. Furthermore, if it is necessary to keep some conventional plant running to allow a fast, local, response to changes on the system, this would tighten the constraint. In other words, there could well be hundreds of hours a year in which it is not possible to export all of the wind output produced in Scotland, or to use it locally. The system operator will have to manage this congestion, constraining off some Scottish wind generators and buying extra power from generators further south.

5. Can BETTA lead to efficient operating decisions?

Given the challenges that renewable generation will pose for the electricity system in Great Britain, will the current trading and transmission arrangements be able to cope? While the set of tasks identified in section 1 ran in the order that decisions are made (from deciding to build capacity, to connecting it to the grid, to operating it), it is the operating decisions that ultimately determine the consequences of, and thus the incentives for, the earlier decisions, and we will consider those first.

How does BETTA fare at the task of ensuring the efficient operation of the stations connected to the network? It relies on the power of arbitrage among the companies trading power. Generators commit much of their capacity through bilateral trades made well in advance of real time, and it is unlikely that these would lead to a cost-minimising generating schedule – or even a feasible one in the presence of transmission constraints. Minimising costs closer to real time depends on higher-cost generators being able to find lower-cost stations able to replace their output and to conclude a trade with them in the short-term markets, which, in Great Britain, trade relatively low volumes of power. Furthermore, the incentive to avoid being penalised for a shortage of power when imbalances are settled still gives generators a reason to try to under-load their stations, increasing production costs.

BETTA's other weakness is in the way in which transmission constraints are managed. National Grid trades in the Balancing Mechanism to ensure that the real-time dispatch does not breach any constraints, but is strongly discouraged from "trading energy", with the implication that it must make as few trades as it can get away with. This means that it has to avoid arbitrage between cheaper and more expensive generators, and does not take transmission losses into account when deciding on a re-dispatch. Green (2007) estimates that the cost of not sending

efficient price signals to generators in England and Wales would have been around 0.1% of their turnover in 1996-7.[5] This is a small proportion, but a large amount in absolute terms – and including Scotland in the market will have raised the level of transmission congestion as the interconnector with England has often been constrained. In the near future, higher levels of renewable generation in Scotland will worsen the problem, and the net present value of congestion between 2010 and 2020 has been estimated at between £1 billion and £3.5 billion (Redpoint, 2010).[6]

It would not be cost-effective to build so much transmission capacity that all constraints were eliminated, and our target should be to use the capacity we have as efficiently as possible. The regulator has previously suggested that rights to use the transmission system should be auctioned to generators and then traded on a short-term basis so that those who were able to make best use of the grid would acquire the right to generate (Ofgem, 2001). The underlying paradigm was that the price of energy should be the same all over the country, but that the price of transmission could vary. Efficient trading of transmission access rights would have produced a price of power, net of the cost of acquiring transmission rights, which varied with the generator's location and sent an accurate signal of the true value of electricity – had the idea been workable. In practice, congestion could only be effectively managed via a very large number of different access rights, since a generator's impact on a constraint can be very sensitive to its location, and markets in the individual rights (some only of interest to a handful of generators) could never be sufficiently liquid for an efficient trading-based solution to occur.

A more radical change would be to acknowledge that in an electrical network, the prices of energy and of transmission cannot in reality be disentangled. The value of transmission between two points is the difference between the value of energy at one of those points and at the other. If the price of energy at each point is set to equal this value – the marginal cost of producing it at points where this is possible, and the marginal cost of producing it somewhere else and moving it to the nodes without generation – then generators will have the correct incentives to increase or reduce output in ways that help the system operator avoid congestion.

[5] The short-term cost of not sending price signals to consumers was much higher, at 1.2% of generators' turnover.

[6] The lower estimate comes from Redpoint consulting, working for the government, while the higher estimate was by Frontier Economics, working for the regulator, Ofgem. According to Redpoint, the spread in estimates is largely due to different assumptions about the amount of renewable capacity built in Scotland, and the speed with which the transmission system is reinforced.

This means that the price of power must be calculated for every point on the system, but there are well-defined algorithms for doing so. They have been used by three markets in the north-eastern United States for more than ten years. The independent system operators in New England, New York and PJM (Pennsylvania, New Jersey and Maryland, but now covering a much wider area) all operate voluntary markets in which generators' offers are used to calculate the marginal cost of power at each point on the network. The day-ahead market accepts bids to buy power and offers to generate and provide reserve capacity for the following day, and the ISO calculates an operating schedule that maximises the net benefit from trading (i.e., the value of accepted bids less the value of accepted offers) while respecting constraints on generators' operations and those on the transmission system. The prices and quantities calculated are "financially firm", committing traders to deliver or accept the amount of electricity involved. If they cannot, a real-time market allows them to unwind those commitments, but at new prices calculated from revised bids and offers from those companies able to adjust their positions at short notice. The markets are voluntary, and much power is traded bilaterally in advance, but the ISO markets are sufficiently liquid to be attractive to generators, ensuring an efficient dispatch.

One disadvantage of these markets is that the so-called locational marginal prices may well be volatile, since they incorporate variations in both the cost of electricity and of transmission. If there is congestion on the transmission system, the price in the exporting area will fall, and the price in the importing area will rise, reflecting the need to rely on relatively expensive local generation. However, that volatility can be hedged, first with the standard forward and futures contracts that guarantee the price of power at a central location, and second with Financial Transmission Rights (FTRs) that lock in the price difference between such a central location and the holder's own site. The power that is moved from an exporting node to an importing one will pay a transmission charge equal to the difference in prices between them, and the revenue from this will go to the transmission system operator. This makes the system operator the natural counter-party to an FTR. Instead of receiving the (volatile) actual price difference when power is moved over the grid, the system operator returns it to the generator, receiving a more predictable income (from initially selling the FTRs) in exchange.

6. Can BETTA lead to efficient connection decisions?

The second task identified in this paper is to ensure that generators can be connected to the system in an efficient manner. In one way, it is easy to assess the performance of the arrangements used so far – they have been tried and found wanting. However, this should lead us to consider the government's proposed alternative, even though it has not yet been implemented. To reiterate, generators will be allowed to connect to the transmission system as soon as any local work has been completed, even if this would worsen transmission constraints elsewhere. The system operator will have to incur the costs of counter-trading to manage these constraints, and will pass them on to all consumers and generators (who are also likely to pass them through to consumers in the power price). Modelling done for the government suggests that this will allow enough generators to connect to the system for the UK to meet its targets for renewable electricity generation, but will also lead to an increase in congestion costs. The predicted increase in these costs is extremely sensitive to the rate at which renewable generation is connected in Scotland (in particular) and at which transmission capacity can be expanded.[7]

The thrust of these reforms is to ensure that generators have no reason to avoid areas that are likely to be transmission constrained.[8] This will not lead to efficient connection decisions. Efficiency requires that generators face the economic consequences of their decisions, and hence that net generation revenues are lower in transmission-constrained areas; either because transmission charges are higher or because electricity prices are lower. In some cases, this may not affect the generator's decision, because the local renewable resource is so good that this outweighs the difference in transmission costs, but if the generator does not face those transmission costs, the question will never be asked.

A system based on nodal pricing would automatically give generators the correct signals about the relative merits of different points of connection – if they choose to locate in an area with frequent constraints, the price they receive will be lower, on average. There are two political problems with this approach. First, it will tend to increase the level of financial support needed by renewable generators (even if it reduces the resource cost of accommodating them). This would either reduce the amount of generation built, or raise the cost of the support scheme.

[7] Redpoint (2010) report a central scenario with an increase in constraint costs of £195 million (in net present value, from 2010 to 2020), whereas Frontier (2009) estimate an increase of £1.7 billion, comparing their central cases.

[8] Generators do currently have an incentive to locate at sites where they may ease congestion, because they will be able to earn more in the Balancing Mechanism for resolving it.

The way to minimise the impact on support costs would be to ensure that the support scheme is very carefully targeted on the generators who really need help, rather than the current mechanisms which offer the same support to all generators using a particular technology, wherever they are sited. In the 1990s, renewable generators had to compete for individual contracts in a series of tender rounds, and this might be a better option for minimising the cost of support.

The second problem is that generators already located in an export-constrained area (or one that is potentially constrained) will see their net revenues fall as capacity is added and the constraint binds more often. One solution to this problem would be to issue long-term Financial Transmission Rights to these generators, allowing them to "lock in" the prices they expected to receive before any entry took place. The seller of these FTRs would expect to lose money, so would have to be compensated, either via a levy on all consumers, or by issuing similar (but profitable) FTRs covering areas where prices were expected to rise in future (and forcing generators currently sited there to accept them). Because FTRs are financial contracts, giving payments whether or not the holder is actually generating power, they would not affect the incentives for efficient operation or plant closure decisions. If the local power prices are insufficient to cover a station's (non-sunk) costs, then the station should close, and the capacity it was using would become available to other generators.

7. Can BETTA lead to efficient capacity decisions?

The first task identified in section 1 was to give generators the incentives to keep the right amount of capacity available to the market. BETTA can be portrayed as an energy-only market, relying on the prospect of high energy prices to incentivise generators to make plant available. In theory, such a market can provide the right incentives, particularly if there is sufficient demand side participation, and prices rise well above the variable costs of the most expensive generators. Most trading under BETTA is bilateral and conducted well in advance, and prices should be higher at times when traders expect the demand on thermal stations to be high. Trading well in advance, however, they will not be able to predict exactly when low wind outputs and high demands create the greatest need for thermal plant. At those times, short-term trades just before Gate Closure and National Grid's actions in the Balancing Mechanism will be needed to keep the system in balance. If generators can see the extreme need for their plant, they

will be able to raise the prices they offer, receiving sufficient revenues. If generators are taken by surprise, they may not be able to adjust their offers to maximise their revenues. In either case, most generators are not going to trade in the Balancing Mechanism, but the incentive properties of an energy-only market require them to receive the same expected revenue. This depends on effective arbitrage between the prices for bilateral trades and those in the Balancing Mechanism and for imbalances. Furthermore, even in the well-organised (and centralised) US markets, prices have not risen to (or stayed at) the levels needed to cover generators' fixed costs, largely because the system operators suppress prices by behaviour such as buying some of their requirements "out of the market" (Joskow, 2008).

In practice, BETTA is not a pure energy-only market. National Grid also buys a significant amount of reserve capacity from generators and large consumers through regular tender rounds (3 of these were held during 2009). Again, arbitrage between these auctions and the other energy markets should ensure that prices in bilateral trades rise to reflect the opportunity cost of not being able to participate in the system operator's tenders, and those generators not selected by National Grid are able to receive an equivalent amount of revenue if they are selling energy.

The problem with this approach is that most generators will only be able to recover their fixed costs if this arbitrage between markets works well, or if market power in the wholesale market raises prices above marginal costs. Relying on market power to ensure sufficient capacity is not an advisable policy, especially as firms can have incentives to keep capacity out of the market in order to reduce reserve margins and bolster their market power (Crampes and Creti, 2005). The combination of an energy-only market, reserve auctions and arbitrage may produce sufficient incentives for the right amount of capacity, but the costs of having too little generation are high.

The regulator is aware of the risks in relying on an energy-only market to remunerate plant that may not run very often, and has considered the issue as part of the so-called Project Discovery, an assessment of the future of Britain's energy markets (Ofgem, 2010). This has concluded that there are significant risks to security of supply in the years approaching 2020, given the amount of investment needed to replace retiring power stations and raise the level of renewable output. Several options for reform have been raised, one of which is to have long-term tenders for low-carbon generation, coupled with shorter-term (perhaps annual) tenders for

conventional capacity. If each station awarded a contract through the long-term tenders received its own tender price, this would allow support to be targeted to the amount actually needed to make that project viable. That would reduce the rents received by generators in favourable locations and make it possible to target transmission costs on the generators causing them without making it more difficult (or expensive) to meet renewables targets.

The short-term tenders could form an annual capacity market, which is another feature of some US power markets (Cramton and Stoft, 2008). They aim to ensure that generators have sufficient incentive to keep capacity available to meet the expected demand, plus reserve needs. Retailers have to buy capacity to cover their own customers' needs in annual auctions held three or four years in advance. The long interval means that entrants as well as incumbents can sensibly bid, since there is time to build a new plant, with several years' payments secured (if an entrant wishes). In return for the payments from the capacity market, generators offer a financial hedge against energy prices, ensuring that they do not gain from both the capacity and the energy markets, and giving them a strong incentive to be available to generate when they are most needed. Ofgem's discussion of the short-term tenders explicitly draws on US experience (2010, pp.45-7), suggesting that it could be an answer to the challenges faced by the UK. The regulator has not yet concluded that capacity tenders are the right answer, and is consulting on both some more limited reforms (finding ways to set higher energy and reserve prices when the market is tight) and on the possibility of setting up a Central Energy Buyer which would decide what capacity should be built. This would be a radical departure from the policies of the past twenty years, and would be premature, given that market-based solutions exist and are working well elsewhere

8. Conclusion

This paper considers the task of integrating a high proportion of intermittent renewable generation into the British electricity market. It concludes that the current rules are inefficient at resolving congestion, may not provide sufficient incentive to make capacity available to meet short-lived peaks in the demand for it, and have not been able to accommodate the queue of entrants wishing to connect to the grid in Scotland. A better approach would be to allow the price of power to vary across the country, reflecting the true state of the transmission system and

giving incentives to reduce generation and investment in constrained areas. This approach has been used for over ten years in a number of US power markets. More recently, some of these markets have developed capacity markets which increase the incentives to ensure that enough generation is available to meet demand. The US markets thus address all of the major problems facing the UK, and provide a "ready-made package" of reforms. Making such significant changes to the British wholesale market would require strong political leadership, but might greatly ease the tasks of creating and operating our future electricity system.

Acknowledgments

This research is funded by the Engineering and Physical Sciences Research Council and our industrial partners, via the Supergen Flexnet Consortium, Grant Number EP/E04011X/1. It has also received support from Advantage West Midlands and the European Regional Development Fund, via the Birmingham Science City Energy Efficiency Project. I would like to thank the editors and two anonymous referees, and the organisers and participants at the IFN and Elforsk Conferences on Market Design, Sweden, September 2009 for helpful comments. The views expressed are mine alone.

References

Bowring, J. (2006) "The PJM Market" in (eds.) F.P. Sioshansi and W. Pfaffenberger (2006) *Electricity Market Reform: An International Perspective*, pp. 451-77, Amsterdam, Elsevier

Crampes, C. and A.Creti (2005) "Capacity Competition in Electricity Markets", *Economia delle fonti di energia e dell'ambiente*, no. 2, pp. 59-83

Cramton, P. and S. Stoft (2008) "Forward Reliability Markets: Less Risk, Less Market Power, More Efficiency", *Utilities Policy*, vol. 16, no. 3, pp. 194-201

Department of Energy and Climate Change (2009a) *Improving Grid Access: Consultation Document*, URN 09D/740 London, Department of Energy and Climate Change

Department of Energy and Climate Change (2009b) *The UK Renewable Energy Strategy*, Cm 7686, London, The Stationery Office

Department of Energy and Climate Change (2009c) *Government Response to the 2009 Consultation on the Renewables Obligation*, Department of Energy and Climate Change, London, URN 09D/847

Frontier Economics (2009) *An assessment of the potential impact on consumers of connect and manage access proposals: A report prepared for Ofgem, November 2009*, London, Frontier Economics

Green, R.J. (2007) "Nodal Pricing of Electricity: How much does it cost to get it wrong?" *Journal of Regulatory Economics*, vol. 31, no.2, pp. 125-149

Green, R.J. and N. Vasilakos (2010) "Market Behaviour with Large Amounts of Intermittent Generation" *Energy Policy*, vol. 38, no. 7, pp. 3211-3220

Gross, R., P. Heptonstall, D. Anderson, T.C. Green, M. Leach and J. Skea (2006) *The Costs and Impacts of Intermittency: An assessment of the evidence on the costs and impacts of intermittent generation on the British electricity network*, London, Imperial College

Hogan, W.W. (1992) "Contract Networks for Electric Power Transmission", *Journal of Regulatory Economics*, vol 4, no 2, September, pp 211-242

Hogan, W.W. (2002) "Electricity Market Restructuring: Reforms of Reforms" *Journal of Regulatory Economics*, Vol. 21, No. 1, pp. 103-132

House of Lords (2008) *The Economics of Renewable Energy, Economic Affairs Select Committee Fourth Report of Session 2007-8*, HL195 of 2007-8, London, The Stationery Office

Joskow, P.L. (2008) "Capacity payments in imperfect electricity markets: Need and design", *Utilities Policy*, vol. 16, pp. 159-170

Newbery, D.M. (1998) "The Regulator's Review of the English Electricity Pool" *Utilities Policy*, vol. 7 no 3, pp.129-142

Ofgem (2001) *Transmission Access and Losses Under NETA: Consultation Document May 2001, 37/01*, London: Office of Gas and Electricity Markets

Ofgem (2010) *Project Discovery - Options for delivering secure and sustainable energy supplies, 16/10* London, Office of Gas and Electricity Markets

Redpoint (2010) *Improving Grid Access: Modelling the Impacts of the Consultation Options A report on a study for the Department of Energy and Climate Change*, URN 10D/549, London, Department of Energy and Climate Change

Wolfram, C. (1999) "Electricity Markets: Should the Rest of the World Adopt the United Kingdom's Reforms?" *Regulation*, Vol. 22 (4): pp.48-53

Appendix – simulated load-duration curves

The load-duration curves in this paper are intended to illustrate the scale of the problems that the electricity industry may face, rather than to provide definitive quantitative predictions. They are based on data used in Green and Vasilakos (2009). The starting point was thirteen years of hourly electricity demand data, from between 1993 and 2005, taken from National Grid's web site. Each year (in turn) was scaled up to a predicted level for 2020, assuming future demand growth of 1.1% a year, and basing the scaling on annual weather-adjusted energy consumption. In other words, hour-to-hour demand variations due to weather conditions are preserved (and will be matched to the wind output that those weather conditions would produce). The procedure also preserves the underlying load shape for each year. As household, industrial and commercial consumption each have their own patterns, and their proportions have changed over our sample period (and are likely to change in future), this is a short-coming of our methodology. However, to correct for this, we would need hourly data on demand by customer type, which are not readily available. The impact of the relative reduction in industrial demand, however, with its (typically) high load factor, will be to make the true 2020 load duration curve peakier than the curves presented here. That would strengthen the arguments in this paper. In the longer term, it is possible that "smart" charging of electric vehicles could flatten the load-duration curve – as long as the market design is capable of sending signals of the best times to do this.

The simulated hourly wind outputs were based on wind speed data from the British Atmospheric Data Centre. Individual weather stations were used to represent 19 onshore and 11 offshore regions, with capacities assigned to them in proportion to the amounts being planned (or built, or existing) in the British Wind Energy Association database. The aggregate outputs were based on 11 GW of onshore capacity and 19 GW offshore. Some missing observations were filled in by interpolation or regression against the wind speeds at nearby weather stations

(including those not used to represent wind regions). The regional output, given the capacity assigned, was based on a standard wind turbine power curve. Some wind speeds were scaled up or down by (individual) constant percentages in order to produce realistic load factors for those sample stations – weather stations are often at sites that would be unsuitable for a wind generator, even if they are in a generally favourable region.[9] Wind speeds from coastal weather stations were also scaled up to obtain suitable average load factors from offshore wind stations. If these average load factors are too high (or too low), then the industry will have to build more (or less) capacity to meet its output targets, implying that any errors in our load factors will be offset in terms of the total output generated – at least on an annual basis. More details on the procedures used are in Green and Vasilakos (2009).

We thus have 13 annual series for demand and for wind generation, each scaled to possible overall totals for consumption and for wind capacity in 2020. Our aggregate load-duration curve for the gross demand, shown in figures 2 and 3, is summed across these 13 years. The net demand on thermal plant (in the same figures) is obtained by subtracting the simulated wind output for each hour from the simulated demand in that hour. Once again, the 13 years of simulated net demands are aggregated to give the load-duration curve. Note that that the point on the gross load-duration curve for the 1,000th highest hour, which corresponds to the 13,000th highest value in the dataset, will be for an hour that is extremely unlikely to also have the 13,000th highest net demand. In other words, points on the two load-duration curves that are vertically aligned will generally correspond to different clock hours, and if the same hours were used for both curves, one would be far from monotonic.

In figure 1, we present the hour-to-hour changes in demand, both gross and net of the (corresponding) change in wind output. In other words, the gross changes series ranks the changes between adjacent hours of the gross demand series described above, while the net changes series ranks the changes between adjacent hours of the net demand series. As before, two vertically aligned points will not (in general) correspond to the same hour.

Finally, figure 4 gives the load-duration curves for Scotland. These are based on four years of data from April 2001, as this was the earliest date for which National Grid publishes detailed Scottish demand data, and our wind dataset ends before a fifth year is completed. The

[9] For example, weather stations at RAF bases typically had the most complete runs of data (which was helpful), but airfields need to be on large areas of flat land, whereas wind farms are often sited on top of hills.

demand data was scaled to possible 2020 levels by the same factors used for figures 1-3, and it was straightforward to split our wind generation series between Scotland and the rest of Great Britain. There was no attempt to calculate different scaling factors for Scottish demand – the aim of the figures in this paper is simply to illustrate the approximate scale of the problems facing the industry in ten years' time.

Figure 1. Hour-to-hour changes in Electricity Demand in Great Britain

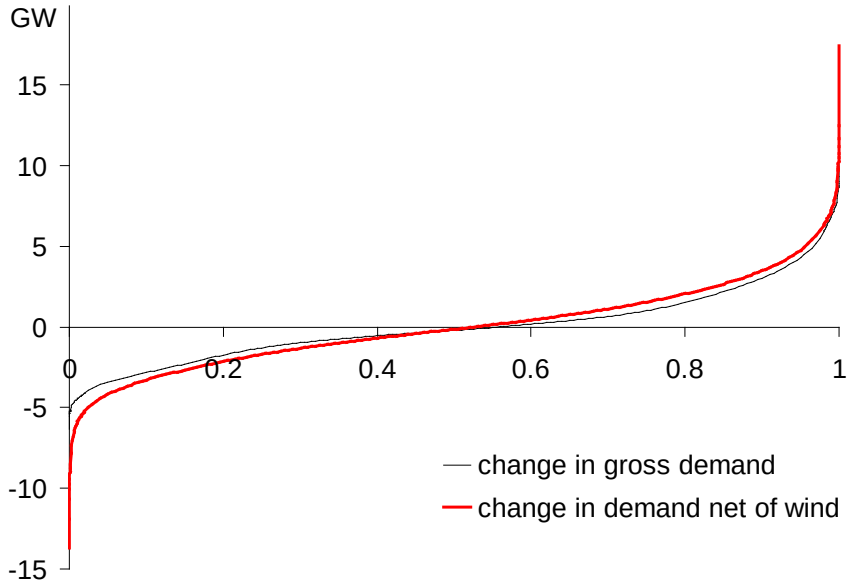

Figure 2. Load-duration curve for Great Britain

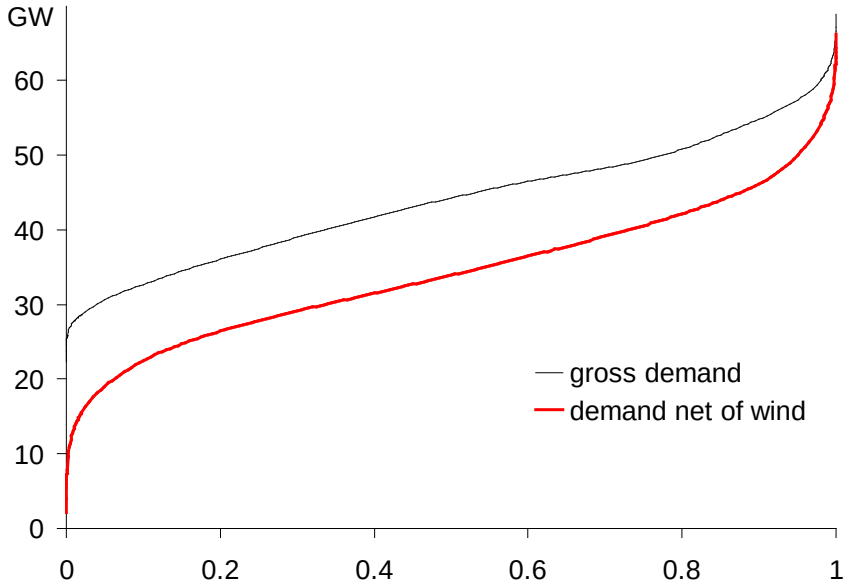

Figure 3. Load-duration curve for Great Britain (detail)

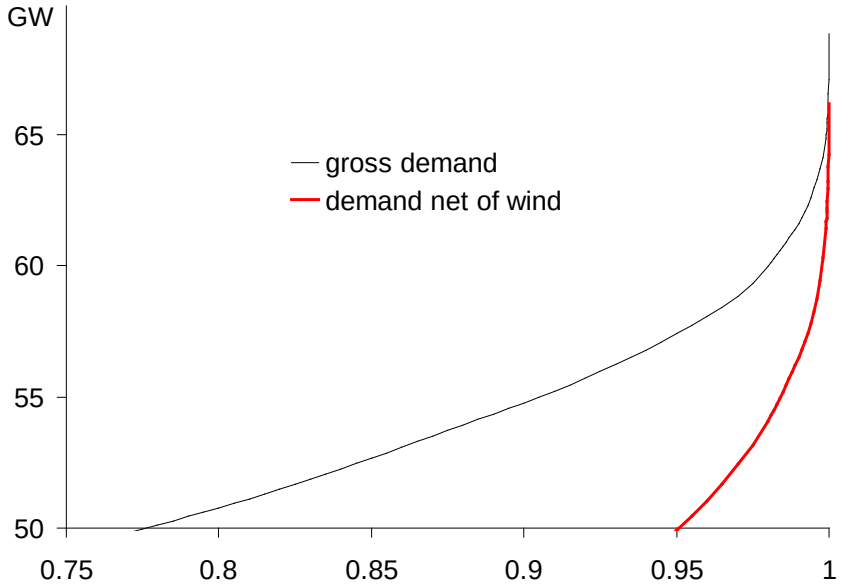

Figure 4. Load-duration curve for Scotland