The Value of Stochastic Programming in Day-Ahead and Intra-day Generation Unit Commitment

Tim Schulze[a,*], Ken McKinnon[a]

[a] *The University of Edinburgh, School of Mathematics, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom*

Abstract

The recent expansion of renewable energy supplies has prompted the development of a variety of efficient stochastic optimization models and solution techniques for hydro-thermal scheduling. However, little has been published about the added value of stochastic models over deterministic ones. In the context of day-ahead and intra-day unit commitment under wind uncertainty, we compare two-stage and multi-stage stochastic models to deterministic ones and quantify their added value. We present a modification of the WILMAR scenario generation technique designed to match the properties of the errors in our wind forecasts, and show that this is needed to make the stochastic approach worthwhile. Our evaluation is done in a rolling horizon fashion over the course of two years, using a 2020 central scheduling model based on the British power system, with transmission constraints and a detailed model of pump storage operation and system-wide reserve and response provision. We show that in day-ahead scheduling the stochastic approach saves 0.3% of generation costs compared to the best deterministic approach, but the savings are less in intra-day scheduling.

Keywords: stochastic programming, unit commitment, hydro-thermal scheduling, wind forecast uncertainty
2010 MSC: 90B05, 90B90

1. Introduction

In recent years the deregulation of energy markets and expansion of volatile renewable energy supplies have led to a significant increase of uncertainty in optimal power systems planning and operation. Several studies have discussed new sources of uncertainty which stem from unpredictable renewable energy supplies: Weber et al [1] developed WILMAR, a stochastic programming model to assess the impact of increased wind power generation on power systems, and Tuohy et al [2] apply this model to test data of the Irish power system. Sturt and Strbac [3] apply stochastic rolling horizon planning to a model of the British power system with a significant amount of wind power, and Constantinescu et al [4] use wind scenarios from a numerical weather prediction model in a two-stage stochastic model. Similarly, Ji et al [5] use two-stage stochastic programming to plan power systems operation under uncertain wind power supply and Falsafi et al [6] investigate the effects of demand response mechanisms in this context. Other studies have identified an increase in traditional uncertain parameters such as load: Nowak and Römisch [7] and Carøe and Schultz [8] both apply stochastic programming to a power system with uncertain demand. These developments have increased interest in stochastic optimization models for short-term power plant scheduling, that is, day-ahead and intra-day generation unit commitment (UC), and a variety of specialised techniques have been developed to speed up the solution of these computationally challenging problems.

Solution and Evaluation Techniques. Many popular algorithms for stochastic unit commitment (SUC) problems are based on decomposition techniques. They can be divided into three groups: Benders decomposition [9], Progressive Hedging [10], and Lagrangian relaxation [11] or Dantzig-Wolfe decomposition [12]. All three approaches are applicable to two-stage or multi-stage models and can be used to decompose the problem by stages, scenarios, or generation units. The different ways of decomposing the problem are reviewed in Römisch and Schultz [13]. Besides the development of decomposition techniques, there have been efforts to accelerate the solution of stochastic problems by bound strengthening through cutting planes: Rajan and Takriti [14] devised facets of the polytope described by minimum up- and downtimes of the generation units and Jiang et al [15] show that these are also facets of the stochastic formulation.

Although substantial efforts have gone into improving solution methods for mixed-integer SUC models, they remain computationally difficult problems. Despite that, comparatively little has been published about the added value of stochastic scheduling models over deterministic ones. In the literature, there are two different approaches to evaluate the expected cost of UC schedules:

*Corresponding author
Email addresses: timschulze@gmx.net (Tim Schulze), k.mckinnon@ed.ac.uk (Ken McKinnon)

1. Evaluation via Monte-Carlo simulation: for the given schedule, a dispatch solution is calculated on a large number of day-long sample paths generated from a simulator that is thought to represent reality. This is typically done for a set of representative days, e.g. one day per season of the year. The performance of different schedules is measured by their expected dispatch cost.

2. Rolling horizon evaluation: a rolling scheduling and dispatch procedure is defined in which the system is scheduled for a few hours and evaluated against a historic trajectory by a dispatch model. Following the evaluation, the next few hours are scheduled and the process is repeated. Performance is measured by the dispatch cost on the historic trajectory. This is sometimes referred to as time domain scheduling simulation.

A major disadvantage of the Monte-Carlo simulation approach is that it is not possible to be certain whether the simulator is a correct representation of reality. Also, intertemporal constraints such as minimum up- and downtimes cannot be considered beyond the end of the simulated day. These shortcomings are avoided in the rolling horizon approach.

Previous Evaluations. The following studies use Monte-Carlo simulation to evaluate UC schedules: Ruiz et al [16] report on an evaluation of deterministic and two-stage SUC under load and generator failure uncertainty, using the IEEE reliability test system [17]. Papavasiliou and Oren [18] apply Lagrangian relaxation and Benders decomposition to solve two-stage stochastic problems with uncertain wind production and security constrained problems with contingency scenarios. They compare different formulations with respect to fuel cost and security of supply by evaluating a typical spring day in the California ISO test system. Constantinescu et al [4] include wind scenarios obtained from a numerical weather prediction model in a two-stage stochastic model. They evaluate this against a deterministic model, using three days of wind data from Illinois and a ten generator test system.

Tuohy et al [2] apply the WILMAR model [1] to data of the Irish electricity system and perform a one year rolling evaluation of deterministic and multi-stage SUC. They report savings between 0.25% and 0.9% when using a stochastic approach instead of a deterministic one, depending on the length of the first stage. However, the authors use perfect information on the first stage, which biases the solutions to become better if the length of the first stage is extended. Additionally, the problems are only solved to an optimality tolerance of 1%. Sturt and Strbac [3] report on the difference between deterministic and stochastic rolling planning in a thermal power system with high wind penetration and a given level of storage capacity, which represents the British power system in 2030. However, mainly continuous relaxations of integer models are used, and transmission network issues arising from the geographical disparity of wind, storage and conventional generation are not addressed.

Our Approach. In this paper we compare the performance of stochastic and deterministic UC approaches in day-ahead and intra-day planning under wind uncertainty, using a two-stage stochastic model in the day-ahead context and a multi-stage stochastic model in the intra-day context. Our study is performed in a rolling horizon fashion over an evaluation period of two years. We use a mixed-integer scheduling model based on the British power system from the perspective of a central scheduler. It includes transmission restrictions between network areas, a detailed pump storage model, and a model of system-wide reserve and response provision. Hence the model can be used to effectively evaluate strategies for dealing with wind forecast errors against the backdrop of the system's flexibility in generation, storage and reserve provision under transmission restrictions. We investigate the fundamental interactions between wind uncertainty, storage and scheduling methods, and these issues are most clearly understood in the setting of a centrally scheduled system. The centrally scheduled situation can provide a reference model when comparing different market structures, but these market issues are not considered in this paper. The system data we use correspond to National Grid's figures for 2020 under the Gone Green Scenario, with a wind penetration of 30% in terms of installed capacity.

While stochastic models are computationally challenging, the savings achieved with these techniques are typically a small percentage of the overall cost, implying the necessity of small optimality tolerances. To solve the problems efficiently to a gap of 0.1%, we use a scenario decomposition approach based on Dantzig-Wolfe decomposition. The method is described in detail in [19]. To generate our scenario trees, we use techniques published in the WILMAR [20] study, however we demonstrate that they need to be adapted to incorporate forecast level dependency of wind forecast errors in order to make the stochastic approach worthwhile.

The remainder of this paper is organised as follows: Section 2 has a formulation of our UC model; Section 3 has details of the input data, scenario generation and scenario tree construction techniques; Section 4 has a description of the rolling horizon evaluation procedure; Section 5 has the evaluation results; and Section 6 has the conclusions.

2. Stochastic Unit Commitment Model

The UC model used for our rolling horizon evaluation includes an aggregated representation of the transmission system with generation zones and transmission links between them. There are limits on the power flow under normal operation. These are expressed in terms of individual transmission links and additional boundaries, each of which splits the network in two and imposes a real power

flow limit on the sum of transmissions crossing it in each direction. The limits are derived by the network operator, using physical network feasibility criteria, N-1 security and contingency analyses [21]. The model contains pump storages which can be used for providing ancillary services and storing wind energy. Each pump storage scheme is modelled as a closed reservoir system, connected to a single plant which contains multiple pump-turbines. Wind power availability is treated as uncertain and a scenario model is used to approximate its possible realisations. Excess wind power can be curtailed at no cost. Load shedding is also permitted, but at a high cost.

In terms of thermal generation units we distinguish fast-start units from slow units. Fast-start units are open-cycle gas turbines (OCGT) which can be started within the hour. All other thermal units are categorised as slow and must be notified at least an hour before they can become available to generate.

Following British practice, we distinguish between frequency response and reserve. Response is fast-acting and is used to stabilise the frequency within seconds, e.g. in the immediate aftermath of a fault, for up to 15 minutes. Reserve is used for two separate reasons: to deal with errors in wind forecasts and to restore response capability by freeing up used response after a failure. Reserve is required to be available for at least an hour. While dedicated variables are needed for frequency response provided by part-loaded generators, reserve can be modelled without additional variables. To do this we formulate one quantity for response, and another quantity for the sum of response and reserve. For pump storage units we use both, dedicated response variables and combined reserve and response variables. Reserve and response are treated as soft constraints, and we include piecewise linear (PWL) functions to penalise for providing insufficient amounts of them. Since the boundary limits were set under contingency considerations, reserve and response are modelled as system-wide services which are not affected by them, i.e. we assume that the boundaries can be overloaded in a post-contingency state where reserve and response are required.

Our day-ahead planning model is a two-stage stochastic model, while the intra-day model is a multi-stage stochastic model. We present a single model here, which can be adapted to represent both situations, depending on the choice of non-anticipativity constraints. A single-scenario version of the same model is used to perform deterministic scheduling and to evaluate existing schedules by solving a dispatch problem.

An overview of our notation is given below, followed by an algebraic model statement. Sets are in calligraphic font, parameters are Latin and Greek capitals, and variables are lower case Latin or Greek letters. Superscripts are used to extend variable names, while subscripts are indices. Reserve and response quantities are distinguished by a hat: for any quantity associated with response, say r, the corresponding quantity for response plus reserve is denoted by $\hat{r}$. The planning horizon is $t = 1, \ldots, T$, and where the statement shows or implies variables for $t \leq 0$, they are fixed input data rather than actual variables.

Sets.

$\mathcal{B}, \mathcal{B}^{01}$:	sets of scenario bundles. Bundles in B^{01} are for binary commitment decisions of slow units.
$\mathcal{D}$:	set of transmission boundaries in the network
$\mathcal{F}$:	set of fast start units with $\mathcal{F} \subset \mathcal{G}$. Slow units in $\mathcal{G} \setminus \mathcal{F}$ require notification prior to startup.
$\mathcal{G}$:	set of generation units, $\mathcal{G}_n$ is the set of generators at node $n \in \mathcal{N}$
$\mathcal{L}$:	set of transmission lines
$\mathcal{N}$:	set of network nodes (transmission areas)
$\mathcal{P}$:	set of pump storage plants, $\mathcal{P}_n$ is the subset at node $n \in \mathcal{N}$
$\mathcal{S}$:	set of wind power scenarios, $\mathcal{S}_b$ is the subset of scenarios associated with bundle $b \in \mathcal{B}$
$\mathcal{W}$:	set of wind farms, $\mathcal{W}_n$ is the subset at node $n \in \mathcal{N}$

Parameters.

Ψ:	minimum proportion of response to be met by part-loaded generators
B_{ld}:	line-boundary adjacency matrix. 1 if line l crosses boundary d in one direction, -1 if it crosses in the other direction, 0 otherwise
$C(r^{tot})$:	PWL penalty function for keeping too little response r^{tot}
$\hat{C}(\hat{r}^{tot})$:	PWL penalty function for keeping too little reserve plus response $\hat{r}^{tot}$
C_q^{nl}:	no-load cost of generator g [\$/h]
$C_q^{H_2O}$:	end-of-day water value in the reservoir of pump storage plant q [\$/MWh]
C_g^m:	marginal cost of generator g [\$/MWh]
C_g^{st}:	startup cost of generator g [\$]
C^{voll}:	value of lost load [\$/MWh]
D^t:	time granularity of the model [h]
D^{res}:	time for which a generator is required to serve response if called upon [h], with $D^{res} < D^t$
E_q:	pump-generator cycle efficiency at storage $q \in \mathcal{P}$ [proportion]
H_q^{max}:	reservoir capacity at pump storage plant $q \in \mathcal{P}$ in MWh of dischargeable energy
N_q^{pum}:	number of (identical) pumps in pump storage plant $q \in \mathcal{P}$
$N_l^{st,end}$:	start (end) nodes of line l
P_q^{cap}:	capacity of a single pump in pump storage plant $q \in \mathcal{P}$ [MWh]
P_{nt}^{dem}:	real power demand at node n in period t [MW]
$P_{g,q}^{min,max}$:	minimum (maximum) generation limit of generator $g \in \mathcal{G}$ (pump storage $q \in \mathcal{P}$) [MW]
$\bar{P}_{l,d}$:	maximum power transmission on line l / across boundary d [MW]
P_s^{rob}:	probability of scenario s
$P_g^{ru,rd}$:	operating ramp up (down) limits of generator g [MW/D^t]
$P_g^{su,sd}$:	startup (shutdown) ramp limits of generator g [MW/D^t]

P^{win}_{wts}: wind power available from wind farm w in period t, scenario s [MW]
R^{max}_g: max response available from generator g [MW]
T: last time period of the planning horizon
T^{nt}_g: startup notification time of generator g [h]
$t^{st,end}_b$: start (end) periods of scenario bundle b
$T^{u,d}_g$: minimum uptime (downtime) of generator g [h]

Variables.

$\alpha_{gts} \in \{0,1\}$: 1 if thermal unit g is on in period t, scenario s, and 0 if it is off
$\gamma_{gts} \in \{0,1\}$: 1 if thermal unit g is started up in period t, scenario s, and 0 otherwise
$\eta_{gts} \in [0,1]$: 1 if thermal unit g is shut down in period t, scenario s, and 0 otherwise
$\delta_{qits} \in \{0,1\}$: 1 if pump i of storage q is pumping in period t, scenario s, 0 otherwise
$\zeta_{qts} \in \{0,1\}$: 1 if storage q is generating in period t, scenario s, and 0 otherwise
$h_{qts} \in [0, H^{max}_q]$: level of storage q *after* period t, scenario s in MWh of dischargeable energy
$p^{dis}_{qts} \in [0, P^{max}_q]$: real power discharged from storage q in period t, scenario s [MW]
$p^{flo}_{lts} \in [-\bar{P}_l, \bar{P}_l]$: real power flow on line l in period t, scenario s [MW]
$p^{gen}_{gts} \in [0, P^{max}_g]$: real power output of generator g in period t, scenario s [MW]
$p^{pum}_{qts} \geq 0$: real power pumped into storage q in period t, scenario s [MW]
$p^{shed}_{nts} \geq 0$: load shed at node n in period t, scenario s [MW]
$r^{gen}_{gts} \in [0, R^{max}_g]$: response provided by generator g in period t, scenario s [MW]
$r^{pum}_{qts} \geq 0$: response provided by pump storage q in period t, scenario s [MW]
$\hat{r}^{pum}_{qts} \geq 0$: reserve plus response provided by pump storage q in period t, scenario s [MW]
$r^{tot}_{ts} \geq 0$: total available response in period t, scenario s [MW]
$\hat{r}^{tot}_{ts} \geq 0$: total available reserve plus response in period t, scenario s [MW]
$u^{win}_{wts} \in [0, P^{win}_{wts}]$: used wind power from farm w in period t, scenario s [MW]

Objective function.

$$\sum_{s\in\mathcal{S}} P^{rob}_s \Bigg[\sum_{t=1}^{T} \sum_{g\in\mathcal{G}} \left(C^{st}_g \gamma_{gts} + D^t C^{nl}_g \alpha_{gts} + D^t C^m_g p_{gts} \right) + \sum_{q\in\mathcal{P}} C^{H_2O}_q \left(h_{q0s} - h_{qTs} \right) + \sum_{t=1}^{T} \left(\sum_{n\in\mathcal{N}} D^t C^{voll} p^{shed}_{nts} + C(r^{tot}_{ts}) + \hat{C}(\hat{r}^{tot}_{ts}) \right) \Bigg]. \tag{1}$$

The objective is to minimise the expected cost of supplying electricity to the economy, including expected losses due to underserved reserve and response and a penalty for lost load. The generation cost consists of startup, no-load and marginal cost terms. These contain fuel and carbon emission costs and a levelised contribution from capital cost, operation and maintenance cost and decommissioning cost [22]. The water level after the last period is treated as variable, and we apply a linear water value to the reservoir level difference created over the course of the planning horizon.

We use penalty functions to model the cost of underserved reserve to the economy. The penalties represent the expected cost of lost load due to generator failure(s) at times where the system lacks sufficient response and reserve to deal with them. The penalty function $C(r^{tot}_{ts})$ models the expected cost of single generator failures at a response level of r^{tot}_{ts}, while $\hat{C}(\hat{r}^{tot}_{ts})$ models the *additional* expected cost of double generator failures at a level $\hat{r}^{tot}_{ts}$ of response plus reserve. To obtain the correct penalty for single and double generator failures, both are applied.

The penalty function for underserved response is calculated as follows. In any time period, consider generator g which is operating at its full capacity, while the amount of available response in the system is x. In case of a failure of this generator, we assume that $D^t \max\{0, P^{max}_g - x\}$ MWh of demand are lost and the system can recover after a time period D^t. If the failure probability in any period is p_g and the value of one unit of lost load is C^{voll}, then the expected total failure cost is

$$C(x) := C^{voll} D^t \sum_{g\in\mathcal{G}} p_g \max\{0, P^{max}_g - x\}. \tag{2}$$

The resulting penalty function is shown in Figure (3a). In our implementation, we approximate the function with seven PWL pieces.

After the failure of a single generator, reserve is used to restore the response level. A subsequent failure in the same period D^t will lead to a loss of load unless the combined amount of response and reserve cover the loss of both generators. Using the same approach as above, we derive an additional penalty function $\hat{C}(y)$ for insufficient levels y of response plus reserve. The cost of lost load due to the failure of a generator tuple (g_1, g_2) is $C^{voll} D^t \max\{0, P^{max}_{g_1} + P^{max}_{g_2} - y\}$ if both generators fail in the same period, and zero otherwise. Let $\hat{\mathcal{G}}$ denote the set of all combinations of generators. Under the assumption that generators fail independently, the expected total cost of double failures while operating at a response plus reserve level y is given by

$$\hat{C}(y) := C^{voll} D^t \sum_{(g_1,g_2)\in\hat{\mathcal{G}}} p_{g_1} p_{g_2} \max\{0, P^{max}_{g_1} + P^{max}_{g_2} - y\}. \tag{3}$$

This is used as additional penalty function for underserved response plus reserve and is shown in Figure (3b). In our implementation, we approximate this function with five PWL pieces.

We assume that all generators fail with equal probability, $p_g = p\ \forall g \in \mathcal{G}$. Further, the formulae assume that all generators are operating at their maximum output level, so the penalties tend to overestimate the expected cost of lost load due to failures. Additional losses due to quick successive failures (before response can be restored) and failures of more than two generators within one hour are not taken into account. However, the cost of single failures is a small percentage of overall cost (cf. Figure 8), and the cost of double failures is an order of magnitude smaller than that (cf. Figure 3). Thus the approximation error can be expected to be small. The minimisation of objective (1) is subject to the following constraints:

Load balance equations. For all $n \in \mathcal{N},\ s \in \mathcal{S},\ t = 1, \ldots, T$:

$$\begin{aligned}\sum_{g\in\mathcal{G}_n} p_{gts} + \sum_{w\in\mathcal{W}_n} u^{win}_{wts} + \sum_{l\in\mathcal{L}:N^{end}_l=n} p^{flo}_{lts} + \sum_{q\in\mathcal{P}_n} p^{dis}_{qts} \\ +p^{shed}_{nts} - P^{dem}_{nt} - \sum_{l\in\mathcal{L}:N^{st}_l=n} p^{flo}_{lts} - \sum_{q\in\mathcal{P}_n} p^{pum}_{qts} = 0.\end{aligned} \tag{4}$$

These ensure that power input and output are equal at all times at all network nodes.

Transmission boundary limits. For all $t = 1, \ldots, T,\ d \in \mathcal{D},\ s \in \mathcal{S}$:

$$-\bar{P}_d \le \sum_{l\in\mathcal{L}} B_{ld} p^{flo}_{lts} \le \bar{P}_d. \tag{5}$$

These inequalities impose restrictions on the transmission across predefined boundaries, by limiting the sum of power flows on lines crossing the boundary in each direction. They are used in addition to the limits on individual power flow variables to model network congestion.

Generator bounds. For all $s \in \mathcal{S},\ t = 1, \ldots, T,\ g \in \mathcal{G}$:

$$p^{gen}_{gts} \ge P^{min}_g \alpha_{gts} \tag{6}$$

$$p^{gen}_{gts} + r^{gen}_{gts} \le P^{max}_g \alpha_{gts}. \tag{7}$$

Constraints (6) and (7) establish the connection between power output, response and on-off variables. When a generator is on ($\alpha_{gts} = 1$), it must generate between the minimum and maximum stable limits, and the response it can provide is limited by its spare headroom (beside the upper limit R^{max}_g). When it is off ($\alpha_{gts} = 0$), the generator's generation and response levels are at zero.

Ramp rate constraints. For all $g \in \mathcal{G},\ s \in \mathcal{S},\ t = 1, \ldots, T$:

$$p^{gen}_{gts} - p^{gen}_{g(t-1)s} \le P^{ru}_g \alpha_{g(t-1)s} + P^{su}_g \gamma_{gts} \tag{8}$$

$$p^{gen}_{g(t-1)s} - p^{gen}_{gts} \le P^{rd}_g \alpha_{gts} + P^{sd}_g \eta_{gts}. \tag{9}$$

Ramp rate constraints (8) limit the *increase* in generation level between two successive periods $t-1$ and t in the case where a generator is on in both periods ($\alpha_{g(t-1)s} = 1$, $\gamma_{gts} = 0$), and in the case where it is started up in the second period ($\alpha_{g(t-1)s} = 0$, $\gamma_{gts} = 1$). Similarly, constraints (9) limit the *decrease* in two successive periods during continuous operation ($\alpha_{gts} = 1$, $\eta_{gts} = 0$) and shutdown ($\alpha_{gts} = 0$, $\eta_{gts} = 1$).

Switching constraints. For all $s \in \mathcal{S},\ t = 1, \ldots, T,\ g \in \mathcal{G}$:

$$\alpha_{gts} - \alpha_{g(t-1)s} = \gamma_{gts} - \eta_{gts} \tag{10}$$

$$1 \ge \gamma_{gts} + \eta_{gts}. \tag{11}$$

These establish the relationship between on-off, startup and shutdown variables. During startup and shutdown procedures they impose a unique solution, however when a generator is continuously on or off for two successive periods, the startup and shutdown variables on the right-hand side may either both be one or both be zero. In the case where a generator is off for successive periods it is always optimal for both, startup and shutdown variables to be zero, as the startup variables have positive cost coefficients in the objective. However, when a generator is on for successive periods, the objective can sometimes be decreased by setting startup and shutdown variables to one, since that relaxes the right-hand side of the ramp constraints (8) and (9). To eliminate solutions where startup and shutdown variables are both one, we include constraints (11). The model with all variables (α, γ, η) is what Ostrowski et al [23] call the 3- Binary Variable Formulation. The integrality restriction of either startup or shutdown variables can be relaxed and it follows from (10) that the relaxed variables must take integer values. We relax shutdown variables, since this gives the best performance with our dataset and solver.

Minimum up- and downtime constraints. For all $s \in \mathcal{S},\ g \in \mathcal{G},\ t = 1, \ldots, T$:

$$\sum_{i=t-T^u_g+1}^{t} \gamma_{gis} \le \alpha_{gts} \tag{12}$$

$$\sum_{i=t-T^d_g+1}^{t} \eta_{gis} \le 1 - \alpha_{gts}. \tag{13}$$

To model minimum up- and downtimes, we use the facet-defining minimum up-down cuts (12) and (13) by Rajan and Takriti [14].

Pump storage operation constraints. For all $q \in \mathcal{P},\ t = 1, \ldots, T,\ s \in \mathcal{S}$:

$$\delta_{q1ts} \le 1 - \zeta_{qts} \tag{14}$$

$$\delta_{q(i+1)ts} \le \delta_{qits} \quad \forall i = 1, \ldots, N^{pum}_q - 1 \tag{15}$$

$$p^{pum}_{qts} = \sum_{i=1}^{N^{pum}_q} \delta_{qits} P^{cap}_q \tag{16}$$

$$\zeta_{qts} P^{min}_q \le p^{dis}_{qts} \le \zeta_{qts} P^{max}_q. \tag{17}$$

Pump storage plants are useful for providing reserve and response, meeting peak demand and storing excess wind power. Binary variables ζ_{qts} determine whether a plant is discharging or not, and constraints (17) link them to continuous discharge variables with lower and upper limits. Within one plant, the pumps all have identical capacities, and they can only be pumping when the plant is not discharging (14). After switching on the first pump, the others are switched on in order from lowest to highest (15) to avoid symmetric solutions. The pumping level is decided by the number of active pumps, since they can only run at full capacity (16). The binaries for each pump, δ_{qits}, could be replaced by a single integer variable $\bar{\delta}_{qts}$ indicating the number of active pumps. However, we use binaries because then δ_{q1ts} can be used in constraints (14), (24) and (25) to indicate whether a plant is pumping or not.

Reservoir constraints. For all $q \in \mathcal{P},\ s \in \mathcal{S},\ t = 1, \ldots, T$:

$$h_{qts} = h_{q(t-1)s} + D^t E_q p^{pum}_{qts} - D^t p^{dis}_{qts}. \tag{18}$$

Reservoir levels are tracked by constraints (18). They are expressed in terms of MWh of electrical energy that would be generated using the contained water. A constant cycle efficiency is applied to incoming energy, thus keeping the model linear by neglecting the head effect which is small. The plants are located at separate sites with no hydrological connection. Also, exogenous inflows are small and lower reservoirs are large, so the water cycle of each pump storage plant is modelled as a single reservoir system with a given storage capacity. This is a good approximation for GB pump storage schemes.

Reserve and response definitions. For $t = 1, \ldots, T,\ s \in \mathcal{S}$:

$$\sum_{g \in \mathcal{G}} \left(\alpha_{gts} P^{max}_g - p_{gts}\right) + \sum_{q \in \mathcal{P}} \hat{r}^{pum}_{qts} = \hat{r}^{tot}_{ts} \tag{19}$$

$$\sum_{g \in \mathcal{G}} r^{gen}_{gts} + \sum_{q \in \mathcal{P}} r^{pum}_{qts} = r^{tot}_{ts} \tag{20}$$

$$\sum_{g \in \mathcal{G}} r^{gen}_{gts} \geq \Psi r^{tot}_{ts}. \tag{21}$$

Equations (20) and (19) define system-wide levels of response and reserve plus response, respectively. Part-loaded generators contribute all spare headroom $P^{max}_g - p_{gts}$ to the slow-acting reserve plus response quantity (19), while, due to ramp limits, they can only contribute a limited amount of their headroom $r^{gen}_{gts} \leq \min\{R^{max}_g, P^{max}_g - p_{gts}\}$ to the fast-acting response quantity (20). The contributions from pump storages are defined in the next paragraph. Constraints (21) require a minimum amount of response to be met by part-loaded generators to avoid relying too much on pump storage units.

Pump storage reserve constraints. For all $q \in \mathcal{P},\ t = 1, \ldots, T,\ s \in \mathcal{S}$:

$$\hat{r}^{pum}_{qts} + p^{dis}_{qts} \leq P^{max}_q + p^{pum}_{qts} \tag{22}$$

$$D^t \hat{r}^{pum}_{qts} + D^t p^{dis}_{qts} \leq h_{q(t-1)s} + D^t p^{pum}_{qts} \tag{23}$$

$$r^{pum}_{qts} + p^{dis}_{qts} \leq p^{pum}_{qts} + P^{max}_q \left(1 - \delta_{q1ts}\right) \tag{24}$$

$$D^{res} r^{pum}_{qts} + D^t p^{dis}_{qts} \leq h_{q(t-1)s} + D^t P^{max}_q \delta_{q1ts}. \tag{25}$$

Pump storage plants can provide different levels of reserve and response, depending on whether they are currently discharging, pumping or spinning in air. Constraints (22) and (23) impose limits on the sum of response and reserve $\hat{r}^{pum}_{qts}$, and constraints (24) and (25) limit the available response r^{pum}_{qts}.

When the plant is in **discharge mode** ($\zeta_{qts} = 1$, $\delta_{q1ts} = 0$), pump variables p^{pum}_{qts} are all zero. Then constraints (22) and (23) state that the current discharge plus reserve and response can exceed neither the maximum discharge nor the remaining energy level in the storage. Constraint (24) states that the response provided during discharge is limited by the headroom available in the turbine, and constraint (25) makes sure that there is sufficient energy stored in the reservoir to meet the discharge during the hour and provide response over a fraction of D^{res} of an hour.

In **pump mode** ($\zeta_{qts} = 0$, $\delta_{q1ts} = 1$) a plant can provide reserve by turning off the pumps *and* starting to discharge. The demand reduction through turning off the pumps is fast enough to meet response standards, while subsequent discharge only qualifies as reserve. The discharge level p^{dis}_{qts} is zero, and constraint (22) limits the provided reserve plus response to be at most the current pumping level plus maximum discharge. Now constraint (23) states that the reserve plus response is bounded above by the amount of energy left in the reservoir plus the current pumping level. Equation (24) says that the available response is upper bounded by the pumping level, while (25) is removed by increasing the right-hand side term by P^{max}_q.

Finally, if the plant has its turbines **spinning in air** ($\zeta_{qts} = 0$, $\delta_{q1ts} = 0$), pump and discharge variables p^{pum}_{qts} and p^{dis}_{qts} are both zero, and reserve plus response is simply bounded above by the maximum discharge (22) and the available energy level (23). The same is true for response and is achieved by equations (24) and (25), only here the energy level contained in the reservoir need only be sufficient to maintain response for a fraction D^{res} of an hour. The energy consumption required to keep the turbines spinning in air is small, so we assume that the plant is always spinning in air when it is not either pumping or discharging.

Non-anticipativity constraints. These determine the structure of the decision tree underlying our optimization model. For binary decisions of slow units we use a specific set of bundles, denoted by $\mathcal{B}^{01}$. The following constraints are included for all $b \in \mathcal{B}^{01},\ j, k \in \mathcal{S}_b : k = j + 1$:

$$\alpha_{gtj} = \alpha_{gtk} \quad \forall g \in \mathcal{G} \setminus \mathcal{F},\ t = t^{st}_b, \ldots, t^{end}_b \tag{26}$$

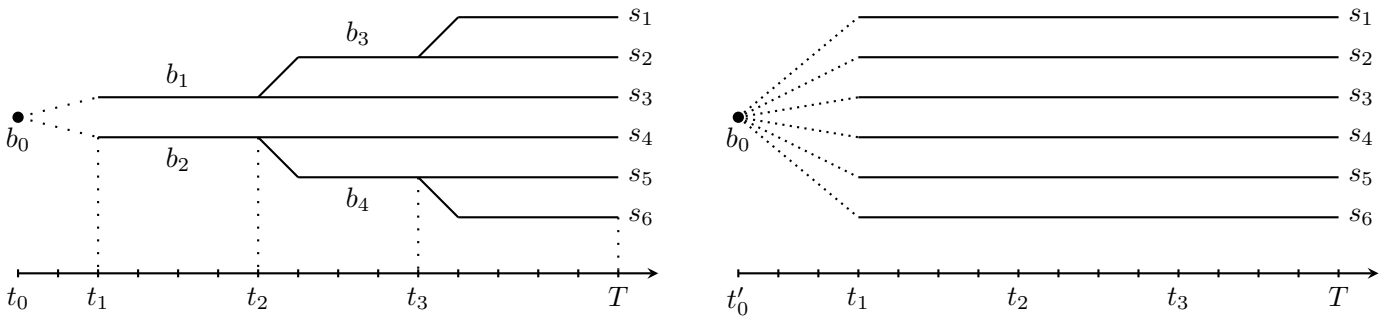


Figure 1: Right: two-stage decision tree with 6 scenarios. First stage commitment decisions are made at time t'_0 and are unique for the whole planning horizon $\{t_1, \ldots, T\}$. From t_1 onwards, recourse decisions in every scenario are made under perfect information. The bundles are $\mathcal{B} = \emptyset$, $\mathcal{B}^{01} = \{b_0\}$ with $\mathcal{S}_{b_0} = \mathcal{S}$ and $t^{st}_{b_0} = 1$, $t^{end}_{b_0} = T$. Constraints (27) are dropped. Left: multi-stage decision tree with two scenarios on the second stage and 6 leaves. The bundles are $\mathcal{B} = \{b_1, \ldots, b_4\}$, $\mathcal{B}^{01} = \{b_0, b_3, b_4\}$. On the first stage we have bundle b_0 with $\mathcal{S}_{b_0} = \mathcal{S}$ and $t^{st}_{b_0} = t_1$, $t^{end}_{b_0} = t_2$. Like the first stage, the second stage also covers periods $\{t_1, \ldots, t_2\}$, with bundles b_1 and b_2 containing $\mathcal{S}_{b_1} = \{s_1, s_2, s_3\}$ and $\mathcal{S}_{b_2} = \{s_4, s_5, s_6\}$. There is a third stage with bundles b_3, covering $\mathcal{S}_{b_3} = \{s_1, s_2\}$ and b_4, covering $\mathcal{S}_{b_3} = \{s_5, s_6\}$, and a fourth stage with no bundles.

$$\gamma_{gtj} = \gamma_{gtk} \quad \forall g \in \mathcal{G} \setminus \mathcal{F},\ t = t^{end}_b + 1, \ldots, t^{end}_b + T^{nt}_g. \tag{27}$$

Constraints (26) make commitment decisions of slow units unique across all bundled scenarios. They are required for both, two-stage and multi-stage stochastic problems.

Constraints (27) are non-standard and are included in multi-stage problems to model startup notification times. When scheduling generators with a deterministic model or a day-ahead stochastic model, a sufficient notification period for generator startups is implicit. However, if commitments are updated in the course of the day, as is done in the multi-stage model, then after a scenario split and decision update we must allow for a minimum notification period to pass before additional startups can become effective. To achieve this, non-anticipativity of startup variables is extended for a notification time after the split of a bundle. During time periods $t^{st}_b, \ldots, t^{end}_b$, constraints (27) are implied by (26) together with (10). Thus, to avoid redundancy we only include them for the time periods $t^{end}_b + 1, \ldots, t^{end}_b + T^{nt}_g$. Further, we use the following additional non-anticipativity constraints for recourse variables of the multi-stage problem. They are included for all $b \in \mathcal{B},\ j, k \in \mathcal{S}_b : k = j + 1$ and $t = t^{st}_b, \ldots, t^{end}_b$:

$$\alpha_{gtj} = \alpha_{gtk} \quad \forall g \in \mathcal{F} \tag{28}$$
$$\delta_{qitj} = \delta_{qitk} \quad \forall q \in \mathcal{P},\ i = 1, \ldots, N^{pum}_q \tag{29}$$
$$\zeta_{qtj} = \zeta_{qtk} \quad \forall q \in \mathcal{P} \tag{30}$$
$$p^{gen}_{gtj} = p^{gen}_{gtk} \quad \forall g \in \mathcal{G} \tag{31}$$
$$p^{dis}_{qtj} = p^{dis}_{qtk} \quad \forall q \in \mathcal{P}. \tag{32}$$

In the rolling horizon evaluation we use deterministic, two-stage stochastic and multi-stage stochastic problems, and with slight data modifications this model represents all of them. Figure 1 shows how we use the data structures to shape two-stage and multi-stage decision trees. The simplest model is a **deterministic** one with a single wind power scenario and no non-anticipativity constraints. It is used for deterministic scheduling and dispatch:

1. In the scheduling model the wind scenario is equal to a central forecast and we use a fixed margin for reserve plus response, i.e. we ask for $\hat{r}^{tot}_{ts}$ to be greater than or equal to some fixed margin.

2. In the dispatch model we fix a given schedule for the slow units and evaluate it against the actual wind outcome. The dispatch model decides optimal output levels of committed generators, operation of fast-start units and pump storage plants, available response and reserve, and the amount of shed load. Thus it compensates for the error in the wind forecast that was used to create the schedule.

The interaction between scheduling and dispatch models in the rolling horizon context is further described in Section 4.

For day-ahead scheduling we use a **two-stage stochastic** model with multiple wind power scenarios as shown in Figure 1 (right). In this setting, the first stage decisions are day-ahead commitments of slow units for the whole 24h planning period. All remaining variables are recourse variables. The two-stage model has non-anticipativity constraints (26), while (27) to (32) are dropped.

For intra-day scheduling we use a **multi-stage stochastic** model as shown in Figure 1 (left). In this model, one stage covers either 3 or 6 hours of the 24-hour planning horizon, depending on how often commitments of slow units can be updated. The first stage decisions are commitments of these units between t_1 and t_2 and startup decisions for a notification time thereafter, which are non-anticipative due to (26) and (27). We use *multiple* wind power scenarios between times t_1 and t_2 to make the commitment decisions robust. This is not standard in the UC literature, where multi-stage trees are typically restricted to a *single* scenario for the first few hours. Within each of the bundles $\mathcal{B} = \{b_1, \ldots, b_4\}$, we seek a non-anticipative

solution by including all constraints (28) to (32). For bundles b_3 and b_4 we also require constraints (26) and (27), while for bundles b_1 and b_2 those constraints are not required because between t_1 and t_2 they are already included for all scenarios, due to bundle b_0. Hence the choice $\mathcal{B}^{01} = \{b_0, b_3, b_4\}$.

We call a schedule non-anticipative when all variables appearing in constraints (8) to (13) and (18) have identical solutions across all subsets of scenarios that are identical at any given time t. Constraints (8) to (13) and (18) are the only constraints which introduce variable interdependence between subsequent time steps, and all variables not appearing therein can be re-evaluated independently at each time step. However, not all variables appearing in these constraints require explicit non-anticipativity constraints if that property can be deduced from other variables linked with them. For pump level variables p^{pum}_{qts}, non-anticipativity follows directly from constraints (16) and (29). Then for reservoir levels it follows from constraints (18), (32) and the fact that initial reservoir levels are fixed. Finally, for startup and shutdown variables it follows from constraints (10) and (11), together with (26) or (28). To avoid unnecessary redundancy, we omit non-anticipativity constraints for those variables for which this property can be deduced from sets of other constraints.

3. Input Data and Scenario Generation

In this section we briefly describe how we obtained data for our model or estimated it where necessary. A graph of the system topology is shown in Figure 2, and the sources of technical system data are described there as well. We outline how historic wind speed forecasts were synthesised, and how forecast error scenarios were generated from a time series model to obtain suitable inputs for the SUC models. Scenario reduction and tree construction methods are also described below. The model uses an hourly resolution, $D^t = 1h$, and we assume that response has to be provided for 15 minutes, $D^{res} = \frac{1}{4}h$.

Lost Load and Unserved Reserve. The GB value of lost load (VOLL) was estimated to be $27,104 (£16,940) per MWh in a publication by London Economics, the Department of Energy and Climate Change (DECC) and the Office of Gas and Electricity Markets (Ofgem) [30]. We use this to model the cost of lost load to the economy and to estimate the penalty functions $C(r^{tot}_{ts})$ and $\hat{C}(\hat{r}^{tot}_{ts})$ for the expected cost of lost load in the case of generator failures. To calculate the cost functions we assume a generator failure probability p which is equivalent to an average of one failure per generator in 150 days. The cost curves are shown in Figure 3.

Synthesising wind power forecasts. Historic wind speed forecasts are not available to us, so for the purpose of this evaluation we synthesise them. Our synthetic forecasts are a weighted average of historic wind and a forecast made

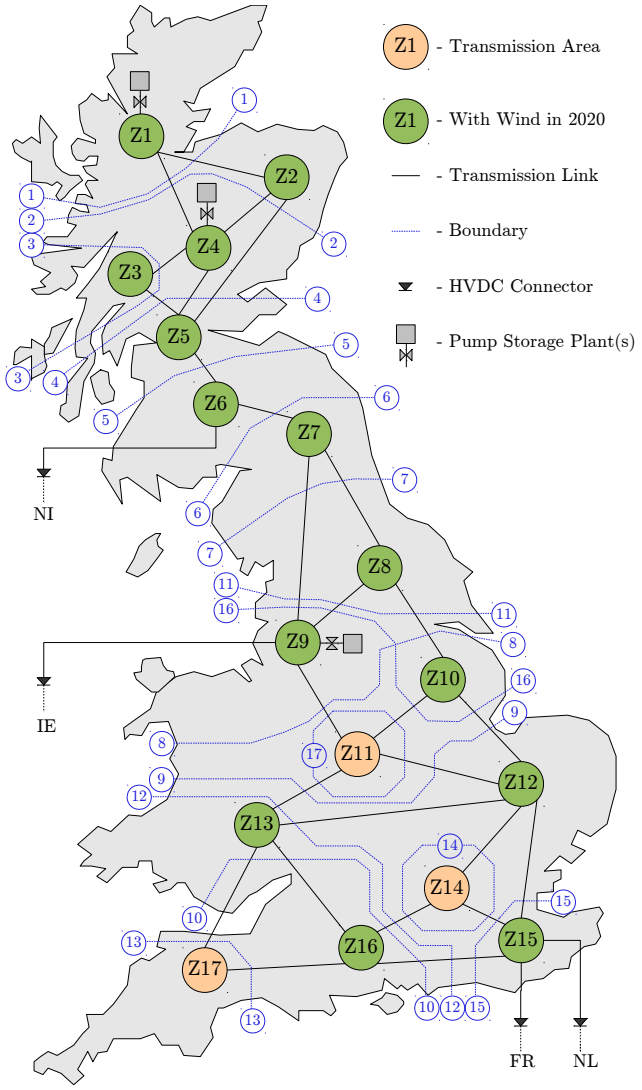
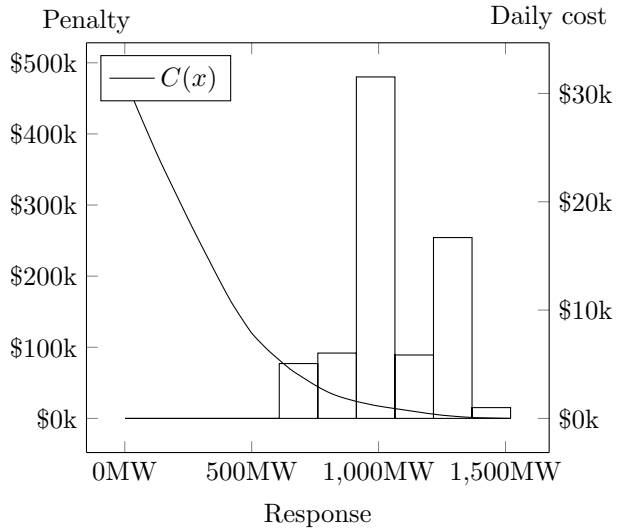
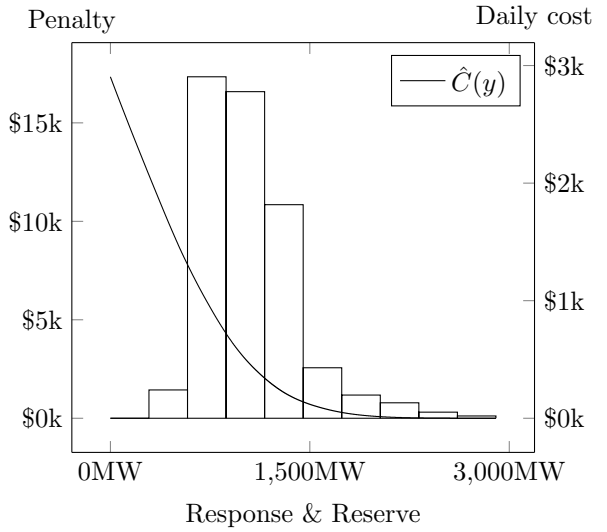


Figure 2: Aggregated GB system with 17 areas and 27 links between them. There are transmission limits on individual links and on boundaries. The 17 boundaries (blue) restrict the sum of transmissions on all lines they cross. There are pump storage plants in zones Z1, Z4 and Z9. Interconnectors to Ireland, France and the Netherlands operate on historic profiles: demand in the affected zones is treated net of interconnector exchange. The data on demand, thermal units, pump storage, wind farms and transmission topology and capacity are from National Grid's 2013 Electricity Ten Year statement [21]. Demand curves are scaled to meet National Grid's 2020 demand expectation. Unit ramp rates and up/down times were obtained through the Balancing Mechanism Report System [24]. Startup notification times are from [25] and [26]. Generation costs were estimated by DECC [22]. They contain carbon cost, fossil fuel cost and a levelised contribution from capital cost and decommissioning cost. Historic response and the proportion Ψ of minimum response from thermal units are calculated from the monthly Balancing Services Summaries, using data for Mandatory Frequency Response and Firm Frequency Response [27]. Historic regional wind speeds are taken from a mesoscale reanalysis [28]. They are translated to load factors with power curves from [29], which are then applied to the wind farms.



(a) Penalty for single generator failures.



(b) Penalty for double generator failures.

Figure 3: Response (a) and reserve plus response (b) penalty functions, based on expected loss due to single and double generator failures. The superimposed bar charts in (a) and (b) show the total cost of operating at different levels of response and reserve plus response, respectively. Discretised in ten bins, the charts show the average daily penalty cost of operating at the given levels of response and reserve. The values are taken from the most successful two-year evaluation of 6-hour deterministic scheduling reported in Section 5.

by pattern matching, in which weights were adapted so as to achieve a root mean square error (RMSE) of 10% of installed capacity at the 24 hour ahead mark. The RMSE matches the shape of typical forecast error curves as shown in Giebel et al [31] and Kariniotakis et al [32]. The graph in Figure (4) shows it as a function of the forecast horizon, along with the RMSE of persistence forecasts and forecasts made by pure pattern matching. Persistence forecasts assume that the current wind conditions will remain unchanged, i.e. persist, and their RMSE is included for reference only. Note that the errors shown here are aggregated over the whole country, which leads to significantly smaller errors than in small regions or single sites [31]. We synthesise forecasts in terms of wind speed and then translate them to regional load factors using power curves from [29]. For the synthesis, we use historic wind speed patterns previous to the year the forecasts are required for. In this context, a pattern is the average of hourly, regional wind speed progressions which satisfy initial criteria for the first two periods. To generate an n hour-ahead forecast, we collect patterns of length $n+2$. Details of the procedure are outlined below.

1. Discretise the domain of observed wind speeds into $k = 1, \ldots, K$ equidistant intervals. Create $3K$ bins $b_{ki},\ i \in \{u, l, d\}$ to hold historic wind progressions $w_j,\ j = 1, \ldots, n+2$. A wind progression is assigned to b_{ku} if the wind in the first hour, w_1 is in interval k and the wind is picking *up*, i.e. $w_2 \geq w_1 + \epsilon$ with a fixed ϵ. Analogously, bins b_{kl} and b_{kd} hold wind progressions that stay *level*, that is, $w_2 \in (w_1 - \epsilon, w_1 + \epsilon)$ or point *down*, so that $w_2 \leq w_1 - \epsilon$, respectively.

2. Iterate over historic wind data. Extract wind progressions of length $n+2$ starting in every hour and assign them to the bins. Find a representative wind pattern $\hat{w}_{jki}$ for every bin b_{ki}, with $j = 1, \ldots, n+2$, by averaging all progressions in the bin.

3. Construct a forecast starting in any hour t with wind history $w_{t-1} \geq w_{t-2} + \epsilon$. If w_{t-2} lies in interval k then the forecast for hours $t, \ldots, t+n-1$ is given by $\hat{w}_{jku}$ with $j = 3, \ldots, n+2$. If the actual wind stayed level or decreased, use forecasts $\hat{w}_{jkl}$ or $\hat{w}_{jkd}$, respectively. Shift the entire forecast $\hat{w}_{jk*},\ j = 3, \ldots, n+2$ by $w_{t-1} - \hat{w}_{2k*}$, so that real wind and forecast are equal in hour $t-1$.

The forecasts are wind patterns which had similar wind speeds in hour $t-2$ and developed similarly in hour $t-1$. They are shifted to match the real wind in hour $t-1$. These pattern forecasts are better than persistence forecasts, but for 6 or more hours ahead they are significantly worse than numerical weather prediction models. Thus, from 6 hours ahead we use a weighted combination of pattern forecasts and actual wind to adapt the RMSE to the level shown in Figure (4).

Generating Scenarios. To generate scenarios, we set up a stochastic model for the time series of the forecast error. The model is fitted to wind forecast errors in the year previous to the evaluation. For instance, to evaluate the scheduling model on 2010 wind data, we collect patterns from 2008 and synthesise forecasts for 2009. Then we use 2009 wind data to calculate forecast errors to which we fit the time series model. The evaluation is then performed on out-of-sample 2010 wind data, with new forecasts generated from 2009 patterns. We evaluate the model on wind

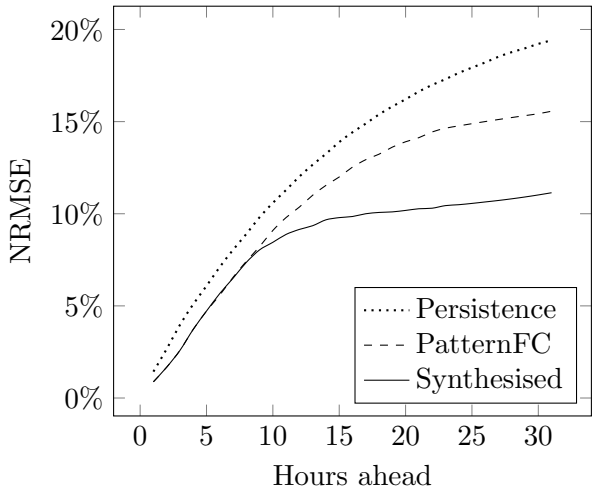


Figure 4: Normalised root mean square error (NRMSE) of wind power persistence forecasts, forecasts made by pattern matching, and our final synthesised forecasts which are a weighted average of pattern forecasts and real wind. Forecast error curves are shown in % of installed capacity and up to 32 hours ahead.

speeds from 2009 and 2010, so the wind data used for this is from 2007 to 2010. The scenario generation technique we use is by Söder [33], and is based on Auto-Regressive Moving Average (ARMA) time series models which are fitted to regional forecast error statistics. The ARMA series contain auto-regressive and moving average terms of lag one, and additional connection parameters are fitted to model the correlation of forecast errors between the regions. In [34] we describe in more detail how we use Söder's approach.

Constructing Scenario Trees. We require multi-stage scenario trees for the intra-day problems and scenario fans for the two-stage day-ahead problems. To generate them, we draw 600 forecast error scenarios from the ARMA model and add them to a synthesised forecast. To keep the stochastic problems tractable, the number of scenarios is reduced with a technique described in Gröwe-Kuska et al [35] and Dupačová et al [36]. Their method finds a subset of the generated scenarios, such that the Kantorovich distance between probability measures of the remaining and initial scenario sets is minimal. This requires a norm that measures the distance between a scenario pair i and j up to time t, for which we use

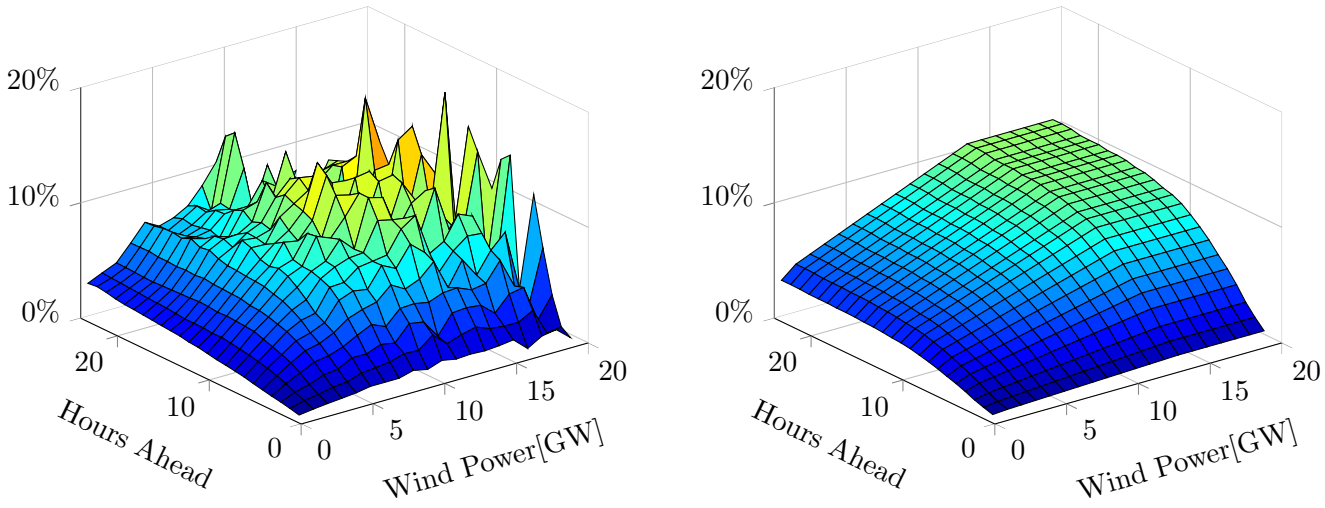
$$d_t\left(i,j\right) := \sum_{w\in\mathcal{W}}\sum_{k=1}^{t}|P^{win}_{wki} - P^{win}_{wkj}|, \tag{33}$$

where P^{win}_{wts} is the wind power from wind farm w at time t in scenario s. We use the selection heuristics described in Römisch et al [35] to find a reduced scenario set which is approximately optimal in the sense of minimal Kantorovich distance and then merge these scenarios into a tree. For the problems in our evaluation we reduce the 600 original scenarios to 12.

Level-Dependent Forecast Errors. The scenario generation methodology described above is based on techniques used in the WILMAR study [20]. The resulting scenarios represent the correlation of wind forecast errors in different regions, but are independent of the forecast wind level. Mauch et al [37] point out that wind power forecast errors are strongly dependent on the forecast levels, so efficient scheduling strategies require wind dependent reserve margins. In order to make stochastic strategies dependent on the forecast level, the variance of the scenario generator must vary with it. Since this is not reflected in the WILMAR scenario generation method, we use a simple scaling approach to adapt the trees used in this study. In each scenario s we replace the original wind at all times t, P^{win}_{wts}, by $(\beta_t P^{win}_{wts} + (1-\beta_t)\bar{P}^{win}_{wt})$ with $\beta_t \in [0,1]$. Here $\bar{P}^{win}_{wt}$ is the average wind under all scenarios and β_t depends linearly on $\bar{P}^{win}_{wt}$. The resulting variance is shown in Figure 5, alongside the root mean square of errors where the forecast overestimated the actual wind. We choose β_t so that the variance of the scenarios matches the RMSE of situations where the actual wind was overestimated because those cases can result in significantly increased cost, due to lost load or the use of expensive fast-start units. On the other hand, cases where the wind was underestimated can be dealt with by curtailing it at no extra cost. The results described in Section 5 show that the scaling leads to a significant cost reduction ($100k per day) in comparison to scheduling with scenario trees which are independent of the forecast level.

4. Rolling Horizon Evaluation

We compare multi-stage stochastic and deterministic scheduling in the intra-day setting, and two-stage stochastic and deterministic scheduling in the day-ahead setting. The evaluation is done in a rolling horizon manner, where 24-hour schedules are made for a central wind forecast or a set of wind scenarios, and then evaluated against the actual wind by solving a set of dispatch problems. After the evaluation step, the planning horizon is moved forward to decide the next schedule. We repeat the procedure until a period of two years is covered and compare the average cost of the different planning techniques. Intra-day UC is performed with 3-hour and 6-hour steps, i.e. the binary decisions for large generators can either be updated every 3 hours or every 6 hours. Day-ahead commitments can only be updated once per day. Other than the update frequency, there are no fundamental differences between the rolling horizon procedures for 3-hour, 6-hour and 24-hour scheduling. In the following, we describe a generic procedure which is applicable to all of them. The process is visualised in Figure 6. The scheduling and dispatch steps are shown separately on the graphic, but are interlaced in the implementation where the rolling procedure alternates between them.



(a) NRMSE of wind power overestimates.

(b) NRMSE of scenario trees.

Figure 5: Left: NRMSE of synthesised forecasts, as function of forecast horizon (hours ahead) and forecast wind level (FC level). Only wind power overestimates were included in the error calculation. Right: NRMSE of the generated scenario trees. The error is scaled in the forecast wind level: for a fixed forecast level above 15GW, any one-dimensional slice through the surface is equal to the NRMSE function shown in Figure 4, while for lower levels the same function is scaled by a linear factor.

Scheduling Steps. Calculating a schedule in practice is not instantaneous and must be done a few hours in advance of its implementation. In Figure 6 (left) we show how a schedule is calculated, using the current system state and a wind forecast. The current state is required to estimate the system state immediately before the implementation of the schedule. After calculating the schedule, it is reported to the dispatch procedure and becomes active a few hours later. We assume that the time between calculating and implementing a schedule is 3, 6 and 8 hours in 3-hour, 6-hour and 24-hour planning, respectively. When implemented, the schedule is active for 3, 6 or 24 hours, and the next one is made in time to become active as soon as its predecessor expires.

Dispatch Steps. The dispatch steps are used to estimate operational costs of implementing a schedule. The dispatch model uses the same hourly granularity as the scheduling model. It is formulated as a 24h problem, but only the first 3 hours are used to estimate the costs. The dispatch model has a single wind scenario with 3 hours corresponding to the actual wind and a forecast for the remainder. Inside the model, the active schedule is fixed, and all recourse decisions are made cost-minimally, that is, the use of OCGT, levels of reserve and response, pump storage operation and shed load. We record the resulting operational cost for a 3-hour period, including penalties for underserved reserve and response and lost load. Additional time periods after the first 3 hours are included to avoid reservoirs being emptied towards the end of 3 hours. The calculated system state is used as initial state for the next dispatch problem. The rolling horizon procedure alternates between scheduling steps and dispatch steps. While schedules are made for 3, 6 or 24 hours at a time, the dispatch model always evaluates 3 hours to keep the results consistent and comparable. Thus, multiple dispatch steps are required after a single scheduling step if the schedule is valid for more than 3 hours. For the dispatch we chose 3-hour steps instead of one-hour steps to save computing time.

Overview of Test Runs. We evaluate seven different types of scheduling: for each of the three approaches with updates every 3, 6 or 24 hours, we run a stochastic and a deterministic version. For reference, we also perform one additional run with perfect foresight, by solving a single-scenario combined scheduling and dispatch model in which future wind power is known in advance. The amount of reserve plus response (19) in this is treated as a soft constraint, unlike in the other deterministic scheduling models which use a fixed margin for reserve plus response.

The stochastic models have the following structure: for 3-hour scheduling we use a multi-stage scenario tree with 3 stages of 3 hours each and a final stage that covers the remainder of the day. There are 3 scenarios on the first stage, then 6, 9 and 12 on subsequent stages. For 6-hour scheduling we use trees which have 4 stages of 6 hours each, with 4 scenarios on the first stage, and then 8, 10, and 12 scenarios on subsequent stages. The two-stage day-ahead model has 12 scenarios on the second stage.

5. Evaluation Results

The results of our two-year evaluation are shown in the graphs in Figure 7. Reserve and response margins for *deterministic* scheduling strategies were set by the formula 'capacity of the largest generator plus r% of the forecast wind'. A range of cases were evaluated with r taking values

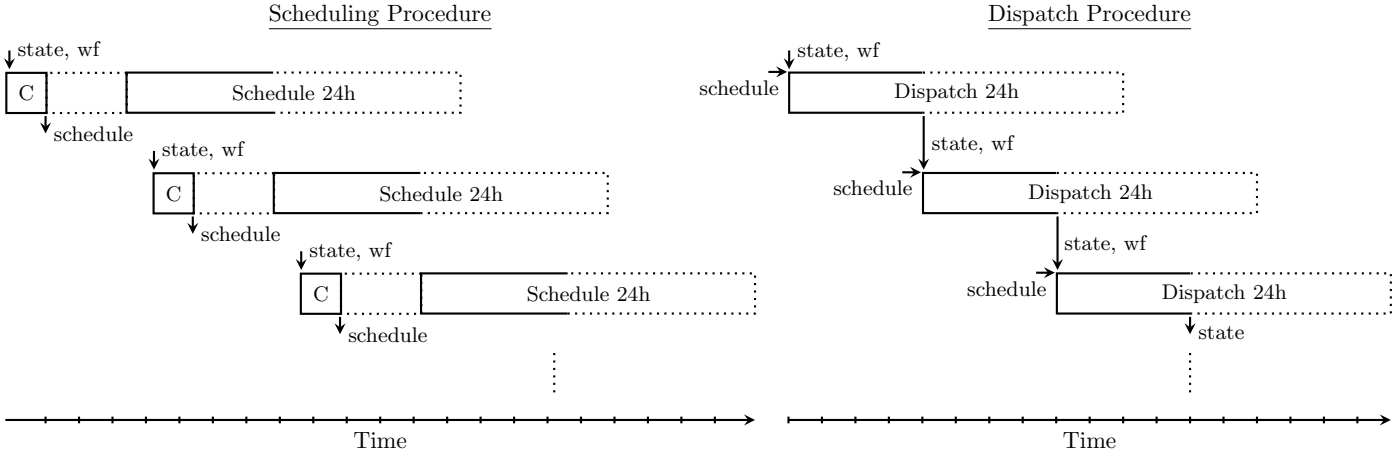


Figure 6: Rolling horizon evaluation procedure. The scheduling steps (left) obtain the current system state (state) and a wind forecast (wf) and calculate (C) a new schedule which becomes active a few hours later (dotted lines) and is valid for 3, 6 or 24 hours (solid part of box). The dispatch steps (right) obtain the current state and the schedule, and evaluate it against the actual wind 3 hours at a time. A wind forecast (wf) is used for the additional 21 hours. The rolling procedure alternates between a scheduling step and (potentially multiple) dispatch steps.

between 0% and 50%, and the resulting average margin is what is shown on the x-axes in Figures 7 and 8. Due to the scenario scaling approach discussed in Section 3, the *stochastic* strategies also depend on wind forecast levels. Stochastic models allocate reserve for forecast errors based on their scenarios, while being aware of the recourse cost of keeping too little additional reserve and response for potential failures. Hence they determine optimal reserve and response levels internally and only need to be evaluated once, unlike the deterministic cases which we evaluated for multiple values of r. On the graphs in Figure 7, the horizontal dotted lines show the values achieved by the stochastic cases with 3-hour, 6-hour and 24-hour schedule updates.

Average Cost. The total cost consists of no-load, startup and marginal generation costs, and various recourse costs. It comes to roughly \$134 per MWh. Recourse costs include the cost of lost load, underserved reserve and response, and OCGT usage. For the 6-hour deterministic cases, the graph in Figure 8 shows a detailed breakdown of these costs. Marginal generation costs of slow units are determined by the demand and the average marginal cost of the committed generators: in Figure 8 they increase from left to right as the amount of OCGT usage decreases and more demand is satisfied from cheaper, slow units. No-load and startup costs of slow units also increase from left to right, while recourse costs decrease from left to right, resulting in cost minima between average reserve and response (R&R) margins of 2.7GW and 3.1GW.

The top left graph in Figure 7 shows an overview of the average daily cost achieved with all scheduling strategies. The deterministic procedures all have cost minima between average R&R margins of 2.5GW and 3GW, where very little or no load is shed and the gradients of increasing no-load, startup and generation costs cancel out with decreasing OCGT and R&R costs. In an area around these minima the cost curve is flat: evaluations with different reserve margins give similar cost. The total cost decreases if commitments of slow units can be revised more regularly: 24-hour scheduling is more expensive than 6-hour scheduling, which in turn is more expensive than 3-hour scheduling. The maximum room for improvement through better forecasts or better (e.g. stochastic) scheduling methods is indicated by the cost under perfect foresight. The average costs with stochastic scheduling models are lower than the minimum costs achieved with the corresponding deterministic models: the gaps are \$100k ($\approx 0.1\%$) per day in the 3-hour and 6-hour cases, and \$300k ($\approx 0.3\%$) per day in the 24-hour case. While the stochastic cases have higher no-load costs than the best deterministic cases, the recourse costs are lower (cf. example in Figure 8).

The cost in the 3-hour stochastic case is \$350k ($\approx 0.35\%$) per day higher than in the perfect foresight case. The gap between these is the value of perfect information. To see if it can be reduced by including more than 12 scenarios in the stochastic model, we performed the same evaluation again with 20 Scenarios. However, the resulting cost did not change significantly ($< 0.01\%$). The stochastic evaluations were also performed without the scenario scaling approach that makes the scenario spread dependent on the forecast level: due to higher no-load and startup cost this led to a worse overall performance of stochastic scheduling, which eliminated the gap between the stochastic procedure and the best deterministic procedure. Achieving minimal operational cost requires a careful balance of committed spare capacity and the use of costly recourse actions, and if the correlation between wind speed and forecast error is not taken into account stochastic models tend to over-commit conventional capacity in situations with low wind.

Load Shedding. The bottom left graph in Figure 7 shows the average annual load shed over the two years. While

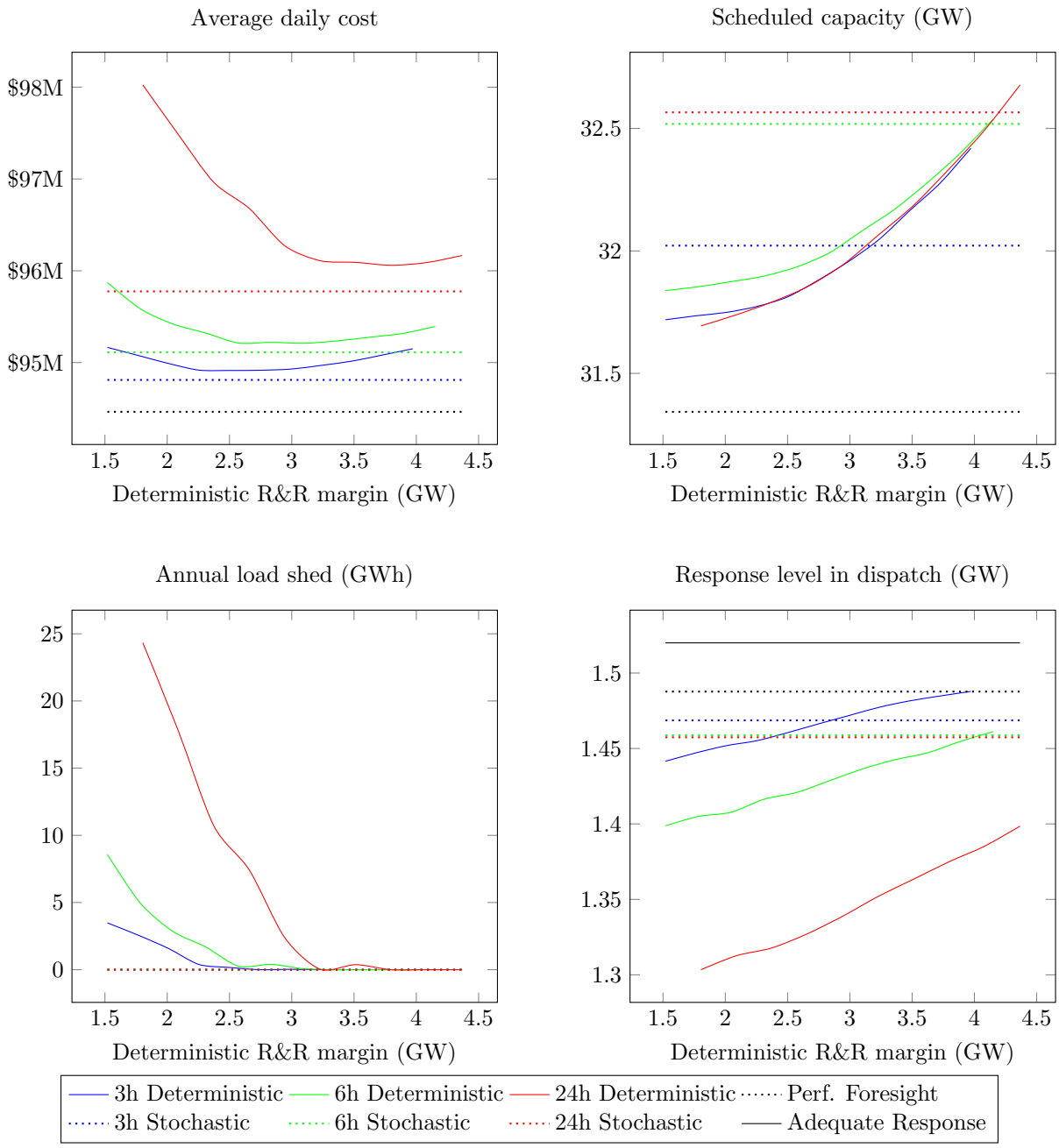


Figure 7: Results of a two-year evaluation of deterministic and stochastic 3-hour, 6-hour and 24-hour scheduling. For reference, the performance of perfect foresight scheduling is also included. The value on the x-axes is the average set margin for reserve plus response (R&R) in deterministic scheduling problems. Stochastic scheduling results are indicated as dotted lines because they are independent of the margin used for deterministic scheduling. The top left graph shows the average daily cost of the different scheduling strategies, including penalties for load shedding and not keeping enough response. The bottom left graph shows the average annual load shed over the two years. The top right graph shows the average conventional generation capacity scheduled by the various approaches. The bottom right graph shows the average amount of response available at the time of dispatch. Here, the 'Adequate' level indicates the level below which a penalty is incurred for not keeping enough response.

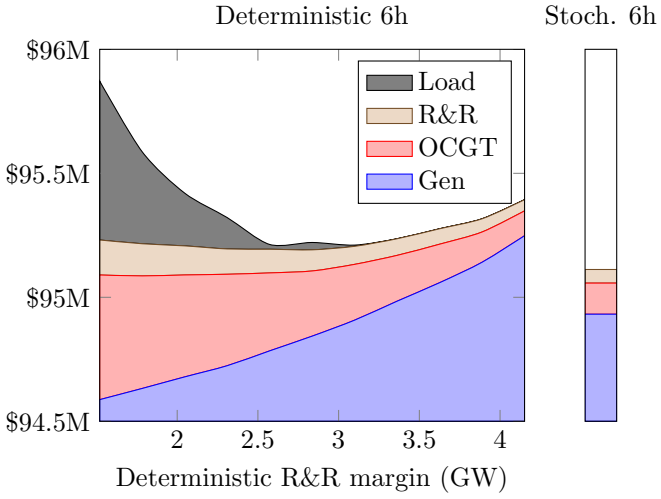


Figure 8: Breakdown of average daily cost into startup, no-load and generation costs of slow units (Gen), and recourse costs. The recourse costs are for OCGT usage (OCGT), underserved reserve and response (R&R) and shed load (Load). The generation cost portion of (Gen) increases from \$91.5M in the leftmost case to \$91.9M in the rightmost case (+\$0.4M). The remaining increase in (Gen) is due to startup and no-load costs which increase from \$3.1M to \$3.34M (+\$0.24M). The graph shows the deterministic 6-hour rolling cases with various fixed reserve and response margins. For the case with average set margin of 2.7GW, Figure 3 shows how the R&R penalty cost is accrued at different levels of underserved reserve and response. For comparison, the average daily costs in the corresponding stochastic case are shown on the bar to the right of the graph.

all stochastic models avoid shedding any load, deterministic models shed load if the set R&R margin is not high enough. The cost of load shedding dominates the shape of the deterministic cost curves in the top left graph of Figure 7. Increasing the deterministic R&R margin does not always lead to a reduction in shed load. In the deterministic model, spare capacity is allocated based on cost only, ignoring potential network congestion, which sometimes leads to situations where it is lumped behind a transmission constraint and unavailable elsewhere. Stochastic models, on the other hand, allocate generation capacity based on correlated wind scenarios and are aware of the network restrictions in the potential operational states. Hence spare capacity is allocated where it is needed to deal with critical wind situations.

Scheduled Capacity. The top right graph in Figure 7 shows the average committed conventional generation capacity. For deterministic cases, this capacity is a consequence of the average wind power level in the forecasts and the set R&R margin. The capacity curves for 3-hour, 6-hour and 24-hour scheduling are all relatively close together. The committed capacity increases with the R&R margin, and this drives the cost increase to the right-hand side of the cost minima in the top left graph.

In stochastic problems, the scheduled capacity is mainly a consequence of the average wind power forecast level and the scenario variation at times for which scheduling decisions are implemented. The relevant forecast horizon is 4 to 6 hours, 7 to 12 hours, and 8 to 32 hours ahead in 3, 6 and 24-hour scheduling, respectively. Figure 5b shows the variation of generated scenarios for these varying forecast horizons: while there is only a small difference between the average errors relevant for 6 and 24-hour scheduling, those relevant for 3-hour scheduling are notably lower. Consequently the 6 and 24-hour rolling procedures committed similar capacity levels, while 3-hour rolling committed a lower level.

System-Wide Response Levels. The bottom right graph in Figure 7 shows the average amount of response in the dispatch. The 'adequate' level indicates where the response penalty curve is first different from zero, and levels below that incur the corresponding penalty. On average, perfect foresight scheduling chooses a level where a small penalty applies, and the stochastic scheduling strategies lead to similar levels. Deterministic procedures lead to lower response levels than their stochastic counterparts, but their level increases with the R&R margin. If we take low response levels as an indicator that the power system is exposed to high stress due to forecast uncertainty, then this shows the stress reduction through stochastic scheduling. The gaps between the deterministic curves and their stochastic counterparts differ systematically: in 3-hour scheduling the curves are at a similar level, while in 6-hour and 24-hour scheduling they are further apart. The higher the forecast uncertainty, the larger the stress reduction through stochastic scheduling. The forecast uncertainty also explains the difference between the top and bottom right graphs: the 6-hour and 24-hour scheduling procedures commit a higher capacity level than the 3-hour procedure, but result in less response at the dispatch stage, as the remaining headroom is used towards dealing with the higher forecast uncertainty.

Network Congestion.. Table 1 shows differences in locational marginal prices (LMPs) between selected zones, averaged over the two-year planning horizon. The LMP values shown here are the dual solutions of constraints (4) for each network zone, taken from the dispatch model. In the presence of transmission restrictions, stochastic models have better awareness of the location where spare capacity is required to deal with forecast uncertainty. The deterministic models are not aware of the spatial correlation of wind forecast errors, which leads to congested situations where neighbouring zones have different LMPs more frequently than with stochastic scheduling. Consequently, the average LMP differences in Table 1 are higher in the deterministic cases. Most cases of congestion appear in Scotland (Z1-Z5), while some also appear in the greater London area and in central England. However, these are less frequent and the average LMP differences are two orders of magnitude lower, so we exclude them from the table. The cases shown in Table 1 include the stochastic case and one selected deterministic case for 3-

Case	Z1–Z2	Z1–Z4	Z2–Z4	Z2–Z5
3hStoch	130.00	156.44	26.46	26.46
3hDet-15	132.57	395.51	262.93	262.93
24hStoch	117.20	132.67	15.47	15.47
24hDet-30	127.31	146.94	19.63	19.63

Table 1: LMP differences between selected zones, averaged over the two-year planning horizon. The LMP values show the average saving per day in $ that can be expected from increasing the transmission capacity between the zones by 1 MW. The shown cases are: stochastic 3-hour scheduling, deterministic 3-hour scheduling with variable R&R margin r=15% of forecast wind (= 2.2GW margin on av.), stochastic 24-hour scheduling, and deterministic 24-hour scheduling with variable R&R margin of r=30% of forecast wind (= 3GW margin on av.).

hour and 24-hour scheduling. As deterministic cases we chose the cost-optimal 24-hour case which incurs no load shedding, and a slightly suboptimal 3-hour case with some load shedding. When load shedding occurs, it drives the LMPs up significantly, resulting in very large LMP differences between zones.

Pump Storage and Congestion Cost. We explore the cost of various model alterations concerning the transmission network and pump storage schemes, by performing evaluations with 3-hour deterministic, stochastic and perfect foresight scheduling. The results are summarised in Figure 9. We show the cost-optimal deterministic strategy, i.e. the case for which the evaluation showed a posteriori that it had the best variable R&R margin. The Z4 case explores the effect of doubling the pump-turbine capability from 440MW to 880MW at the Cruachan storage scheme in zone Z4; the Z1 case explores the effect of adding two storage schemes identical to Foyers in zone Z1 which has 300MW pump-turbine capability and 6.3GWh storage capacity; the NoN case removes all network restrictions; and the NoS case removes all existing pump storage capability from the system.

Taking the stochastic results as the base, the gap between stochastic and perfect foresight solutions is between 0.35% and 0.37% ($350k to $370k daily) in all considered cases. The gap between the deterministic and stochastic solutions is 0.11%, 0.08%, 0.06%, 0.08% and 0.14% in the Norm, Z4, Z1, NoN and NoS cases, respectively (left to right in Figure 9). The additional storage capabilities in the Z4 and Z1 cases reduce the gap between stochastic and deterministic planning from the Norm case, while removing storage capacity in the NoS case increases the gap. Storage provides a way of compensating for wind forecast uncertainty, so it reduces the advantage of stochastic scheduling over deterministic scheduling.

With the best implementable (stochastic) policy, the cost savings achieved by storage expansion Z4 in comparison to the Norm case is 0.05% ($50k), while storage expansion Z1 gives a 0.08% ($80k) improvement. Removing network congestion (NoN) has a value of 0.05% ($50k),

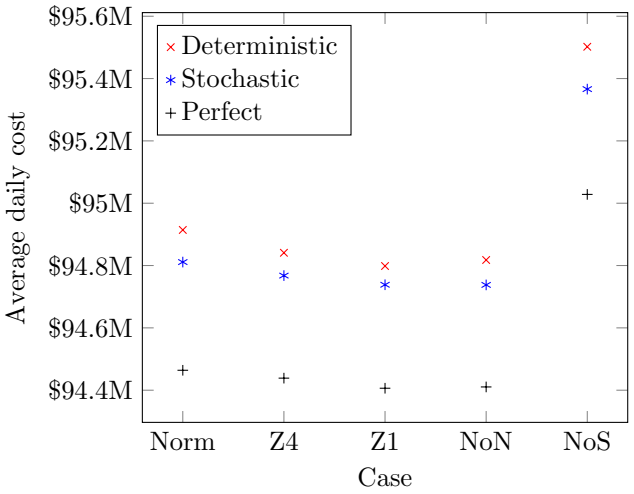


Figure 9: A comparison of operational costs in various cases. The graph shows the average daily cost of 3-hour scheduling with the best (a posteriori) deterministic strategy, stochastic strategy, and perfect foresight. In order from left to right the cases are: the 'normal' reference case (Norm, same as Figure 7), a case with doubled pump-turbine capability but unchanged storage capacity in zone 4 (Z4), a case with two additional pump storage schemes identical to Foyers in zone 1 (Z1) (increased pump, discharge and storage capacity), a case without transmission network restrictions (NoN) and a case without any pump storages (NoS).

while removing storage capabilities entirely (NoS) leads to a major cost increase of 0.6% ($600k). The different cost in the studied cases can be explained with the system's ability to use more wind and commit less thermal generation: in the Norm case 0.36% of available wind power are curtailed, while in the Z1 and NoN cases only 0.05% are curtailed, and in the NoS case 0.7% need to be curtailed. In the Z1 and NoN cases the total capacity of committed thermal generators is on average 0.1% lower than in the Norm case, while in the NoS case it is 2.5% higher. Pump storages provide a major share of system-wide reserve and response, and without them more thermal generators must be switched on and kept off their upper limits to provide reserve and response. The Z4 case does not differ much from the Norm case in terms of wind curtailment, but has 0.06% lower thermal commitment.

6. Conclusion

We have performed a two-year rolling horizon evaluation of stochastic and deterministic unit commitment approaches under wind uncertainty, with periods of varying length between times when the schedules of slow generators can be revised. For the evaluation we use a central scheduling and dispatch model based on the British power system under National Grid's Gone Green scenario for 2020, including a pump storage model and transmission restrictions between network areas. The focus of our study is on the performance comparison of deterministic and stochastic generator scheduling at different time scales. We quantify the monetary value of stochastic scheduling

models over deterministic ones under a central scheduling hypothesis, and pinpoint other advantages of stochastic schedules. We find that

- There are significant cost differences between operating systems that allow major generators to be rescheduled every 3, 6 or 24 hours (all subject to notification times).
- Stochastic models result in minimum operational costs without having to tune reserve margins in advance. In all cases there is a gap between the lowest deterministic cost and the cost of stochastic scheduling. This is despite the fact that we compare their performance with the *best* (a posteriori) setting for deterministic reserve and response which is not known a priori an can be different from one year to another. The superiority of stochastic scheduling grows with the amount of uncertainty in the relevant wind forecasts, but is reduced if additional pump storage capacity is installed.
- We use a simple scaling approach to make wind power scenario trees dependent on the forecast level and show that this leads to a better balance of committed spare capacity and the use of costly recourse actions. This indicates that it may be worthwhile exploring other forecast level dependent scenario generation techniques, e.g. the logit transformation approach by Mauch et al [37].
- Penalties for keeping too little response and reserve are modelled, and they account for \$50k to \$350k ($\approx$0.05% to 0.15%) of the total daily cost. The main cost drivers are generation costs and the recourse cost for shed load and OCGT usage. Stochastic models tune committed capacity levels internally, and the resulting response levels under wind uncertainty are similar to those achieved with a perfect foresight model.
- In the presence of transmission restrictions, stochastic models have a better awareness of spatial wind forecast error correlation, so they place reserve where it is required. Network congestion is measured by differences in LMPs, and it is shown that these are less with stochastic scheduling models. The average cost of network congestion is \$40k per day lower with stochastic scheduling than with deterministic scheduling.

Acknowledgements

We would like to thank Dan Eager and Ian Pope from AF Mercados EMI in Edinburgh for their help with obtaining data for the GB model, Peter Kelen from PowerOP for providing feedback on the model, Samuel Hawkins and Gareth Harrison from the School of Engineering at Edinburgh University for providing the wind data, and Paul Plumptre (former National Grid) for explaining National Grid's balancing mechanism and modelling approach. The first author acknowledges funding through the Principal's Career Development Scholarship scheme of the University of Edinburgh.

References

[1] C. Weber, P. Meibom, R. Barth, H. Brand, WILMAR: A stochastic programming tool to analyze the large-scale integration of wind energy, in: Optimization in the Energy Industry, Springer, 2009, pp. 437–458.

[2] A. Tuohy, P. Meibom, E. Denny, M. OMalley, Unit commitment for systems with signicant wind penetration, IEEE Transactions on Power Systems 24 (2009) 592–601.

[3] A. Sturt, G. Strbac, Efficient stochastic scheduling for simulation of wind-integrated power systems, IEEE Transactions on Power Systems 27 (1) (2012) 323 – 334.

[4] E. Constantinescu, V. Zavala, M. Rocklin, S. Lee, M. Anitescu, Unit commitment with wind power generation: Integrating wind forecast uncertainty and stochastic programming, Tech. Rep. ANL/MCS-TM-309, Argonne National Laboratory (2009).

[5] B. Ji, X. Yuan, Z. Chen, H. Tian, Improved gravitational search algorithm for unit commitment considering uncertainty of wind power, Energy 67 (1) (2014) 52–62.

[6] H. Falsafi, A. Zakariazadeh, S. Jadid, The role of demand response in single and multi-objective wind-thermal generation scheduling: A stochastic programming, Energy 64 (1) (2014) 853–867.

[7] M. P. Nowak, W. Römisch, Stochastic lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty, Annals of Operations Research 100 (1-4) (2000) 251–272.

[8] C. C. Carøe, R. Schultz, A two-stage stochastic program for unit commitment under uncertainty in a hydro-thermal power system, in: Preprint SC 98-11, Konrad-Zuse-Zentrum für Informationstechnik, 1998, pp. 98–113.

[9] Q. P. Zheng, J. Wang, P. M. Pardalos, Y. Guan, A decomposition approach to the two-stage stochastic unit commitment problem, Annals of Operations Research 210 (2013) 387–410.

[10] S. Takriti, J. R. Birge, E. Long, A stochastic model for the unit commitment problem, IEEE Transactions on Power Systems 11 (3) (1996) 1497–1508.

[11] N. Gröwe-Kuska, W. Römisch, Stochastic Unit Commitment in Hydro-thermal Power Production Planning, Preprints aus dem Institut für Mathematik, Humboldt Universität Berlin, 2002, Ch. 30, pp. 605–624.

[12] T. Shiina, J. R. Birge, Stochastic unit commitment problem, International Transactions in Operational Research 11 (2004) 19–32.

[13] W. Römisch, R. Schultz, Multistage stochastic integer programs: An introduction, in: Online Optimization of Large Scale Systems, Springer, 2001, pp. 581–600.

[14] D. Rajan, S. Takriti, Minimum up/down polytopes of the unit commitment problem with start-up costs, Tech. Rep. RC23628 (W0506-050), IBM Research Division (2005).

[15] R. Jiang, Y. Guan, J.-P. Watson, Cutting planes for the multi-stage stochastic unit commitment problem, Tech. Rep. SAND2012-9093J, Sandia National Laboratories (2012).

[16] P. A. Ruiz, C. R. Philbrick, E. Zak, K. W. Cheung, P. W. Sauer, Uncertainty management in the unit commitment problem, IEEE Transactions on Power Systems 24 (2009) 642–651.

[17] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Ij, R. Mukerij, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, C. Singh, The IEEE reliability test system - 1996, IEEE Transactions on Power Systems 14 (1999) 1010–1020.

[18] A. Papavasiliou, S. S. Oren, A comparative study of stochastic unit commitment and security-constrained unit commitment using high performance computing, in: 2013 European Control Conference (ECC), 2013, pp. 2507–2512.
[19] T. Schulze, A. Grothey, K. I. M. McKinnon, A stabilised scenario decomposition algorithm applied to stochastic unit commitment problems, Tech. Rep. ERGO 15-009, The University of Edinburgh, School of Mathematics (2015).
[20] R. Barth, L. Söder, C. Weber, H. Brand, D. J. Swider, WILMAR deliverable 6.2 (d) - methodology of the scenario tree tool, Tech. rep., Institute of Energy Economics and the Rational Use of Energy, University of Stuttgart (2006).
[21] National Grid plc, 2013 electricity ten year statement (ETYS), `www2.nationalgrid.com/UK/` (November 2012).
[22] Department of Energy & Climate Change (DECC), Electricity generation costs 2013, `www.gov.uk/government/publications/electricity-generation-costs` (July 2013).
[23] J. Ostrowski, M. F. Anjos, A. Vannelli, Tight mixed integer linear programming formulations for the unit commitment problem, IEEE Transactions on Power Systems 27 (1) (2012) 39–46.
[24] Elexon Ltd, Balancing mechanism reporting system (bmrs), `www.bmreports.com/` (October 2012).
[25] L. Balling, Flexible future for combined cycle, Modern Power Systems 30 (12) (2010) 61–63.
[26] L. Balling, Fast cycling and rapid start-up: new generation of plants achieves impressive results, Modern Power Systems 31 (1) (2011) 35–40.
[27] National Grid plc, Monthly balancing services summaries (MBSS) 2013/14, `www2.nationalgrid.com/UK/` (2013/14).
[28] S. L. Hawkins, High resolution reanalysis of wind speeds over the british isles for wind energy integration, Ph.D. thesis, The University of Edinburgh – School of Engineering (November 2012).
[29] European Wind Energy Association (EWEA), Integrating wind – developing europes power market for the large-scale integration of wind power, `http://www.trade-wind.eu/` (May 2009).
[30] London Economics, The value of lost load (voll) for electricity in great britain – final report for OFGEM and DECC, `londoneconomics.co.uk/publications/` (July 2013).
[31] G. Giebel, P. Sørensen, H. Holttinen, Forecast error of aggregated wind power, Tech. Rep. Risø-I-2567(EN), Risø National Laboratory (April 2007).
[32] G. Kariniotakis, P. Pinson, N. Siebert, G. Giebel, R. Barthelmie, The state of the art in short-term prediction of windpower – from an offshore perspective, in: Proceedings of the 2004 SeaTech Week, 2004.
[33] L. Söder, Simulation of wind speed forecast errors for operation planning of multi-area power systems, in: International Conference on Probabilistic Methods Applied to Power Systems, IEEE, 2004, pp. 723–728.
[34] T. Schulze, Stochastic programming for hydro-thermal unit commitment, Ph.D. thesis, The School of Mathematics, The University of Edinburgh (2015).
[35] N. Gröwe-Kuska, H. Heitsch, W. Römisch, Scenario reduction and scenario tree construction for power management problems, in: 2003 IEEE Bologna PowerTech Conference Proceedings, Vol. 3, 2003.
[36] J. Dupačová, N. Gröwe-Kuska, W. Römisch, Scenario reduction in stochastic programming: An approach using probability metrics, Mathematical Programming 95 (3) (2003) 493–511.
[37] B. Mauch, J. Apt, P. M. S. Carvalho, M. J. Small, An effective method for modeling wind power forecast uncertainty, Energy Systems 4 (4) (2013) 393–417.