A Risk Function for the Stochastic Modeling of Electric Capacity Expansion

A. Marín,[1] J. Salmerón[2]

[1] *Departamento de Matemática Aplicada y Estadística, ETSI Aeronáuticos,*
Universidad Politécnica de Madrid, 28040 Madrid, Spain

[2] *Operations Research Department, Naval Postgraduate School,*
Monterey, California 93943

Received July 1997; revised April 2001; accepted 25 April 2001

Abstract: We present a stochastic optimization model for planning capacity expansion under capacity deterioration and demand uncertainty. The paper focuses on the electric sector, although the methodology can be used in other applications. The goals of the model are deciding which energy types must be installed, and when. Another goal is providing an initial generation plan for short periods of the planning horizon that might be adequately modified in real time assuming penalties in the operation cost. Uncertainty is modeled under the assumption that the demand is a random vector. The cost of the risk associated with decisions that may need some tuning in the future is included in the objective function. The proposed scheme to solve the nonlinear stochastic optimization model is Generalized Benders' decomposition. We also exploit the Benders' subproblem structure to solve it efficiently. Computational results for moderate-size problems are presented along with comparison to a general-purpose nonlinear optimization package. © 2001 John Wiley & Sons, Inc. Naval Research Logistics 48: 662–683, 2001

Keywords: capacity expansion; power generation planning; generalized Benders' Decomposition; stochastic programming

1. INTRODUCTION

Electric utilities make long-term capacity-expansion decisions to face customers' future demands. An optimal expansion plan should guarantee a certain quality for the service at the lowest cost for the utility, but, since future demand is uncertain when these decisions are made, a utility cannot know the actual impact of a given plan beforehand.

Our previous work [21] presents a similar model to the one proposed in this paper. That work compares performances for different decomposition schemes. The present work provides insight on the practical implications of the model, details specific features involved in the formulation, the treatment of uncertainty, and compares them to other approaches in the literature. A new feature added to the model is the piecewise representation of generation costs. In addition, a comprehensive description of the use of generalized Benders' decomposition to solve the problem is included.

Correspondence to: A. Marín

© 2001 John Wiley & Sons, Inc.

Hobbs [17] provides an excellent survey on the range of models that have been developed for electric utility resource planning and identifies specific areas that require further research: demand uncertainty, risk, and a number of different objectives, among others.

Uncertainty modeling can be conducted through different strategies which may depend on available historical data and forecasting capabilities. Also, decision makers play an important role in the modeling of a problem, ensuring that the results meet their needs and practical requirements. A well-known text on this topic is Wets [36]. A more recent review on modeling and decomposition methodologies applied to large-scale stochastic problems is Ruszczynski [31].

One of the first applications of stochastic programming to electric capacity expansion modeling is Dapkus and Bowe [9]. This work employs stochastic dynamic programming methodology (e.g., Winston [37]) by considering a number of stages (time periods), and states (the result of combining technologies, their availability, and the peak level of demand). Despite the good results achieved with this model, its performance is strongly dependent on the number of states, which in turn complicates the ability to consider additional uncertain data, or to consider independent plants (or generation units) instead of employing generic technology types.

Scenario-based approaches plan against a set of given scenarios which represent a number of possible realizations of the uncertain parameters. These approaches have raised serious concerns about the tradeoff between the proposed models accuracy and their dimensions since problem complexity increases as the number of scenarios grows. These models usually involve a large number of control and decision variables to represent the ‘‘implementable policies’’ (decisions independent of the actual scenario to occur).

Some important works on applied scenario modeling are the following: Gorenstein et al. [15] present an innovative approach for power-system planning in terms of minimization of regret. This is being applied in many others fields involving optimization under uncertainty. Álvarez et al. [2] use a scenario framework to characterize the uncertainty on the right-hand side (rhs), minimizing the regret of wrong decisions and forming implementable policies. Also, Berman, Ganz, and Wagner [4] introduce a scenario-modeling approach for capacity-expansion planning in the service industry.

In the last decade parallel computing has provided researchers with a method to handle a large number of decision variables. The capabilities of new high-performance computing platforms are being used for two purposes: (a) parallelizing classical algorithms (such as Cholesky's factorization for interior point algorithms and branch and bound for integer programming, e.g., Birge and Holmes [5] and Yang and Zenios [38]) and (b) implementing typical decomposition schemes (such as Benders' decomposition or Lagrangean relaxation–decomposition), which usually lead to scenario-separable subproblems. We do not intend performing any parallel computations, but we note that our decomposition procedure is amenable to parallelization as in some of the works described next.

Ariyawansa and Hudson [1] and Ermoliev and Wets [11] have assessed the parallel performances of Dantzig-Wolfe, Benders, and general nested decomposition schemes. Also, Hiller and Eckstein [16] explore a model for managing asset/liability portfolios with stochastic cash flows by a massive parallel Benders' cuts implementation.

Another work that exploits a scenario framework is Mulvey and Ruszczynski [23]. The authors decompose a general multistage stochastic problem into scenario subproblems that can be solved independently on a distributed network. The decomposition scheme in this case is based upon an augmented Lagrangean scheme which dualizes the nonanticipativity constraints and performs a diagonal quadratic approximation. One of the test cases in that research is a dynamic financial planning model described in Mulvey and Vladimirou [22]. Several financial problems with a number of stochastic parameters are analyzed, and the authors compensate for the large size of the model by exploiting its network structure. The nonanticipativity condition is implemented under a multistage scheme where the stages correspond to time periods at which new information becomes available. The result is a significant acceleration in computation. Another work in the financial field is Golun et al. [14], which presents a multiperiod stochastic model that is empirically superior to immunized portfolios and dynamic single-period stochastic models. Ruszczynski [30] presents a parallel decomposition of multistage problems, incorporating regularized quadratic terms in subproblems that preserve the finite termination property of classical approaches. His computational experience shows that the advantages of parallelizations increase with the size of the scenario tree. Also, Vladimirou [35] adapts the regularized and linear decomposition on distributed multiprocessors using a set of standard stochastic programs, showing significant improvements in problems with numerous first-stage variables.

Escudero et al. [12] describe a two-stage model for hydrothermal coordination in power generation planning under uncertainty, implementing a decomposition scheme based on splitting variables for each scenario related to the last deterministic (and implementable) time period. The proposed framework is suitable for a parallel implementation.

Robust optimization is a relatively new technique for dealing with uncertainty. Rockafellar and Wets [29] identify robust decision policies by discovering similarities among the optimal solutions for different scenarios. A general framework for robust optimization is analyzed in Malcolm and Zenios [20] and Mulvey, Vanderbrei, and Zenios [24]. These works assess different choices for parameters that penalize the variance of the cost and the norm of infeasibilities. A serious concern is to strike the proper balance among the terms that form such an objective function: On the one hand, the tradeoff between the mean and variance of the solution provided by the model; on the other, the infeasibility factor due to deviations from a feasible solution for all the scenarios. The specific application of this technique to capacity-expansion models aims to attain an ‘‘almost’’ optimal solution for realizing any of the demand scenarios (robustness) while preserving a reduced excess capacity for any of these realizations.

We particularly note an independent development to our own. Malcom and Anandalingam [19], which describes a comprehensive real-world application of robust optimization to electric capacity expansion, including multiple supply and demand regions, a great variety of physical and resource restrictions and environmental features. They also show how robust optimization leads to consistent choices, both ‘‘solution robust’’ and ‘‘model robust’’: The optimal solution from the model is near-optimal for any realization of the demand scenarios (solution robust) and has ‘‘almost’’ no error regarding uncertain parameters (model robust). Sensitivity to input data and comparison to classical stochastic programming are analyzed.

Referring to other optimization techniques under uncertainty, stochastic dual-dynamic programming was introduced in Pereira and Pinto [27]. The aim of the work is to introduce coupling operational variables between consecutive periods (e.g., water storage in hydro reservoirs) without discretizing them into a number of selected states. This avoids the combinatorial explosion of states through a multistage decomposition framework that takes advantage of the interpretation of the dual solutions as Benders' cuts. With some of these ideas, the aforementioned work [15] also deals with the stochastic planning of electric systems.

We also mention the following important papers on electric utility expansion models, even though they do not consider uncertainty in demand: In Murphy and Wang [25], a multiperiod capacity-expansion model is formulated as a network model, Sherali and Staschus [33] consider the possibility of using renewable energies. The authors explicitly use a load duration curve to formulate a nonlinear objective function. The model is solved by Benders' decomposition and Lagrangean relaxation techniques. The only stochastic extension proposed by the authors consists of simulating a security condition by increasing the rhs coefficients associated to the demand.

A real implementation in Mexico, represented by an integer problem and including renewable energies, can be found in Sherali, Staschus, and Huacuz [34]. Also, Nooan and Giglio [26] formulate a large-scale nonlinear mixed-integer model for planning electric capacity expansion. The authors emphasize the reliability of electricity supply. A combination of successive linearization and Benders' partitioning leads the authors to very attractive computational results, implemented for a real-world problem.

Other remarkable classical works on long-term expansion planning are Bloom [7] and Bloom, Caramanis, and Charny [8], who refer to the generalized Benders' decomposition model implemented in the EGEAS program. In particular, [8] improves the procedures in [7] for representing the probability distributions of unserved energy by using the Gram–Charlier series. It also modifies the production cost subproblem for multiple units of standard size. The electric system capacity is given by the sum of (random) individual plant or unit capacities. These works focus on assuring system reliability by imposing a fixed limit on the amount of expected unserved energy, assuming that the load is known in advance. Obviously, the required computation to obtain the aforementioned probability distribution (or, equivalently, the probability distribution for the system power capacity) including all the generation units is rather intensive.

Referring to capacity deterioration, an important paper, including a general formulation for deterministic capacity expansion, is Rajagopalan [28]. This author suggests relating the available capacity to the facility age and use.

Our own paper is organized as follows: In Section 2 the mathematical model is formulated: Decision variables, parameters, and the penalty function of the model are defined. A discussion to relate our model to other models in stochastic programming is included. In Section 3, we explain how to use a Benders' decomposition-based methodology to solve the problem. To solve the Benders' subproblem efficiently, its structure is exploited and the related Kuhn-Tucker optimality conditions are explicitly solved. Other considerations for solving the master problem are commented on. Section 4 provides computational results for small to moderate-size problems. Finally, Section 5 explains our final conclusions.

2. ELECTRIC CAPACITY EXPANSION: FRAMEWORK

In this section we describe the Electric Capacity Expansion Model (ECEM) and discuss the key ideas behind our representation of demand uncertainty.

Our approach assumes an intensive statistical preprocessing to derive a probability distribution representation of the demand for each time subperiod (which consists of a number of hours, not necessarily consecutive, with a similar load). The proposed model combines some of the techniques of stochastic optimization [22, 23, 29, 36], robust optimization [19, 20, 24, 29], and statistical decision rules (De Groot [10]).

The objective function is formed by the following terms: capital investment cost (by unit of expanded capacity), operating costs (represented by a convex piecewise linear function for each energy type or plant), and an additional term to represent the risk associated with future outcomes. The risk function is defined as the expected cost of adjusting the system operation to actual data, once this information is available. For each subperiod a different risk strategy may be adopted.

The model does not require us to assume any specific value for the demand. Instead we employ the cost of modifying an existing generation plan once the demand information becomes available over time. The (*a priori*) marginal probability distributions for the demand are model inputs as well.

Our approach leads to a large nonlinear optimization model. In order to solve this model, decomposition techniques are required. Among them, the generalized Benders' scheme (Geoffrion [13]) has provided us with the best results, so that is the scheme we use in this paper. This methodology has the advantage of yielding a separable subproblem (over time) that can be explicitly solved by analyzing its optimality conditions. This structure appears suitable for future extensions of the formulation (e.g., to represent the capacity-expansion decisions via integer variables).

In this section, we first introduce the general mathematical formulation of the model, then describe the objective function and feasibility constraints, and finally we discuss the model and clarify its practical implementation.

2. 1. Mathematical Formulation: Introduction

We will use the following indices:

i: Plants, for $i = 1, \dots, I$,
t: Periods, for $t = 1, \dots, T$ (e.g., years),
s: Subperiods of period t, for $s = 1, \dots, S^t$ (e.g., blocks of the load duration curve).

Let L_{ts} be the length (hours) of subperiod s at time period t. Let also $L_t = \sum_{s \in S^t} L_{ts}$ be the length of period t.

The decision variables are:

x_{it}: *Power capacity installed* at plant i in time period t (MW),
z_{its}: *Expected generation* at plant i in time subperiod s of period t (MWh).

Notice that the term $\sum_{s \in S^t} (z_{its}/L_t)$ represents the average generation at plant i during time period t.

Let b_i^0 be the initial capacity at plant i (if any) (MW). The total *available capacity*, $\tilde{x}_{it}$, at plant i in time period t is defined as

$$\begin{cases} \tilde{x}_{i1} = x_{i1} + b_i^0, \\ \tilde{x}_{it} = \alpha_i \frac{\sum_{s=1}^{S^{t-1}} z_{i,t-1,s}}{L_{t-1}} + \beta_i \left(\tilde{x}_{i,t-1} - \frac{\sum_{s=1}^{S^{t-1}} z_{i,t-1,s}}{L_{t-1}} \right) + x_{it}, \qquad \text{for } t \geq 2, \end{cases} \tag{1}$$

where $0 < \alpha_i \leq \beta_i \leq 1$. The coefficient $1 - \alpha_i$ represents the *deterioration capacity rate* at (the end of) a given time period, per unit of energy generated in plant i. $1 - \beta_i$ is the same for the nonused capacity. Hence, this approach assumes that deterioration is a function of hours of use, as well as age. Even if the deterioration rate is slow, the average available capacity during the life cycle of a given generation unit decreases with age, due not only to physical loss of capacity, but also more frequent maintenance requirements and downtimes, until finally the plant becomes obsolete.

By an easy inductive argument, we can derive from (1) that

$$\tilde{x}_{it} = \left(\sum_{t' \leq t} \beta_i^{t-t'} x_{it'} + \beta_i^{t-1} b_i^0 + \sum_{t' \leq t-1} \beta_i^{t-t'-1} (\alpha_i - \beta_i) \frac{1}{L_{t'}} \sum_{s \in S^{t'}} z_{it's} \right). \tag{2}$$

2. 2. Objective Function

The ECEM objective function can be stated as:

$$f(x,z) = \text{Investment Costs} + \text{Oper.\&Maint. Costs} + \text{Risk} = C_{IOM} + \text{Risk}.$$

C_{IOM} is composed of the (first and second) linear terms. Define the following (present value) costs:

IC_{it}: Capital (investment) cost for plant i in time period t ($/MW),
OC_{its}: Operating cost of plant i in subperiod s of period t ($/MWh),
MC_{it}: Maintenance cost of installed capacity at plant i in time period t ($/MW),
MCU_{it}: Maintenance cost of plant i due to use in time period t ($/MWh).

The operating costs are typically convex functions of the generated energy, and include fuel and labor costs. The fuel costs are strongly related to the efficiency of generation. This can be approximated by a suitable piecewise-linear function (representing the so-called incremental heat rates of the generators, $k = 1, \ldots, K_i$ for plant i). Therefore, the operating costs for a plant i at time (t,s) can be stated as:

$$\sum_{k=1}^{K_i} OC_{its,k} z_{its,k}, \qquad \text{for } OC_{its,k} \leq OC_{its,k+1}, \qquad \text{and} \qquad 0 \leq z_{its,k} \leq \bar{z}_{its,k},$$

where $z_{its} = \sum_{k=1}^{K_i} z_{its,k}$ and $\bar{z}_{its,k}$ is the length of the efficiency range k. For the sake of simplicity, we assume a fixed linear cost OC_{its} (to be modified by the term above if needed), but will explain in Section 3 how the methodology is valid for the piecewise-linear generalization too.

C_{IOM} can be stated as

$$C_{IOM}(x,z) = \sum_i \sum_t \left[IC_{it} x_{it} + MC_{it} \tilde{x}_{it} + \sum_{s=1}^{S^t} (OC_{its} + MCU_{it}) z_{its} \right].$$

Since $\tilde{x}_{it}$ is defined in terms of (x,z), by eq. (2), we can express the above terms as

$$C_{IOM} = \sum_i \sum_t Q^x_{it} x_{it} + \sum_i \sum_t \sum_s Q^z_{its} z_{its} = Q^x \cdot x + Q^z \cdot z.$$

The risk term is represented by a separable nonlinear function $R(z)$:

$$R(z) = \sum_t \sum_s W_{ts}(\tau_{ts}), \qquad \text{where} \quad \tau_{ts} = \sum_i z_{its}. \tag{3}$$

Each function $W_{ts}(\tau_{ts})$ represents the expected penalty due to the need to adjust an existing generation plan to new conditions (e.g., changes in demand), assuming that this information is not available at the moment of making the system-expansion decisions. τ_{ts} is a decision variable that represents the energy (MWh) that we expect to supply during the subperiod s of period t.

We assume that there exists a random variable D_{ts} representing the forecast load (MWh) to be served in subperiod s of period t.

In order to evaluate risk, we define the following function:

$P_{ts}(\tau_{ts}, d_{ts}) =$ penalty function ($) to adjust the initial generation plan τ_{ts} to the true demand d_{ts}.

It is usually accepted that P_{ts} depends on $(d_{ts} - \tau_{ts})$ through a function $g_{ts}(\cdot)$ as follows:

$$P_{ts}(\tau_{ts}, d_{ts}) = \begin{cases} g_{ts}(d_{ts} - \tau_{ts}), & \text{if } d_{ts} > \tau_{ts}, \\ 0, & \text{if } 0 < d_{ts} \leq \tau_{ts}. \end{cases}$$

As an instance, we can take $g_{ts}(u) = k_1 u^{k_2}$ for given $k_1 > 0, k_2 \geq 1$ for any real $u \geq 0$.

The function $g_{ts}(u)$ is especially important for the peak subperiods, where the load excess may become unserved energy and imply major regret. This point is discussed in more detail in Section 2.4.

The function P_{ts} must satisfy some general conditions; for example, it must be nonnegative, increase as $(d_{ts} - \tau_{ts})$ grows, be convex in τ_{ts} for each fixed d_{ts}, and satisfy certain boundary conditions.

Finally, we define $W_{ts}(\tau_{ts})$ as the expected penalty for this time subperiod. Let $f_{ts}(d)$ be the density function associated with the random variable D_{ts}, and take

$$W_{ts}(\tau_{ts}) = \int_{\Re^+} P_{ts}(\tau_{ts}, v) f_{ts}(v)\, dv, \qquad \forall \tau_{ts} \geq 0 \tag{4}$$

$W_{ts}(\tau_{ts})$ allows us to evaluate the expected impact of a given decision beforehand. (4) becomes a finite sum of terms if we consider a discrete probability distribution for demand.

The criterion assumed in the formulation of the risk function (minimization of the expected regret) is appropriate under risk neutrality (Wald criterion). Other strategies based upon utility functions could be adopted, although the methodology described in this paper requires the convexity of such functions (see [10]). For details on some of the strategies to hedge against uncertain outcomes, we refer again to the works [14], [20], and [22]. An extensive analysis on risk aversion from a statistical point of view may also be found in Luce and Raifa [18].

In this paper, besides the objective function itself, risk aversion is incorporated by imposing a minimum amount of energy to be supplied in each time subperiod (e.g., a given percent of the maximum demand; see Section 2.3). The joint effect of cost minimization and minimum supply constraints allows us to compromise between the desired profitability and the quality of the service (possibly according to additional regulations). We expect that the minimum supply constraints are binding when the penalty functions for expected unserved demand are not costly enough to compensate for the expense of larger investments. Birge and Louveaux [6] describe a similar idea taken from nonlinear utility models.

Another classic strategy for hedging against risk that can be analyzed with our formulation is Savage's minimax criterion; see [10]. To do this we assume that the decision maker provides, in advance, an appropriate degree of confidence $\epsilon \ll 1$. A large value m is associated with the most unfavorable demand load such that $\Pr\{D > m\} = \epsilon$. Since the objective function (once the demand is known as $D = d$) would have the form

$$f(x, z/D = d) = \sum_i Q_i^x x_i + \sum_i Q_i^z z_i + P\left(\sum_i z_i, d\right),$$

where we have dropped the (t, s) subindices for simplicity. Under an ϵ-minimax strategy we would have

$$\min_{x,z} \max_{d} f(x, z/D = d) = \min_{x,z} \sum_i Q_i^x x_i + \sum_i Q_i^z z_i + P\left(\sum_i z_i, m\right),$$

given that the loss P is ϵ-maximum at the value $d = m$. Since P is convex by hypothesis, the methodology used for solving the Benders' subproblem remains valid; see Section 3 for details. We also note that this criterion may be easily implemented with only a modest amount of data because the probability distribution D is not required. Instead, a unique value is used to represent the (maximum) demand that must be met.

2. 3. Physical Constraints

Constraints of ECEM are defined as follows:

$$z_{its} \leq a_{its} L_{ts} \tilde{x}_{it}, \qquad \forall i, t, s, \tag{5}$$

$$l_{ts} \leq \sum_i z_{its} \leq m_{ts}, \qquad \forall t, s, \tag{6}$$

$$\sum_i IC_{it} x_{it} \leq g_t, \qquad \forall t, \tag{7}$$

$$\sum_t x_{it} \leq h_i, \qquad \forall i, \tag{8}$$

$$0 \leq x_{it} \leq k_{it}, \qquad \forall i, t, \tag{9}$$

$$z_{its} \geq 0, \qquad \forall i, t, s, \tag{10}$$

where

a_{its} = availability factor for plant i in subperiod s of time period t $(0 \leq a_{its} \leq 1)$,
l_{ts} = security level for the energy to be served in subperiod s of time period t (MWh) (e.g., l_{ts} = mean of D_{ts}),
m_{ts} = upper extreme of the support of D_{ts} (MWh) ($+\infty$ is allowed),
g_t = budget for capacity investment in time period t (\$),
h_i = maximum capacity to install at plant i over the time horizon (MW),
k_{it} = maximum capacity to install at plant i in time period t (MW).

Note: The availability factor a_{its} is the expected fraction of the time for which plant i is available for use. This parameter can be estimated, assuming that the random failures of each plant follow a renewal process, so that a forced outage occurs with probability $1 - a_{its} = \bar{r}_{its}/(\bar{m}_{its} + \bar{r}_{its})$, where $\bar{m}_{its}$ is the so-called "mean time between failures" and $\bar{r}_{its}$ is the "mean time to repair."

Finally, ECEM can be expressed as

$$\text{ECEM}: \qquad \min_{x,z} f(x,z) = Q^x \cdot x + Q^z \cdot z + R(z),$$

$$\text{s.t.} \begin{cases} Bx + Az \leq b^0, \\ x \in \mathcal{X}, \\ z \in \mathcal{Z}, \end{cases}$$

where B and A are the matrices that represent constraints (5), $\mathcal{X}$ is the feasible set defined by (7)–(9), and $\mathcal{Z}$ represents (6) and (10).

2. 4. Model Discussion

The proposed ECEM may be considered as a stochastic programming model where the risk function plays the role of the second-stage penalties to guarantee solution robustness. In the discrete (or scenario-based) case, a typical equation for demand satisfaction for a single subperiod would have the form

$$\sum_i z_i + x_2(d_\omega) = d_\omega, \qquad \text{for each scenario } \omega \in \Omega, \text{ with } \omega \text{ finite}.$$

Here $x_2(d_\omega)$ would be a second-stage control variable conveniently penalized as discussed below. However, if a continuous representation of the demand D is assumed (as in the real case) one can be sure that $\Pr\{\exists \omega \in \Omega \text{ with } D = d_w\} = 0$. Hence, the control variables are in fact pseudocontrol variables that concentrate an infinite range of possible deviations (from the true demand) at a single mass point.

In the problem being addressed, we are implicitly considering as control variable

$$x_2(d) = d - \sum_i z_i, \qquad \forall d \in \text{support of } D.$$

Obviously, we cannot work directly with an infinite number of constraints. Therefore, we have considered it appropriate to ‘‘dualize’’ these constraints, giving them an interpretation as risk represented by $E_D[P(x_2(d))]$.

From a stochastic-programming point of view, ECEM may suggest that both x and z are first stage variables, so the model might be interpreted as a simple recourse problem. However, by considering the control variables $x_2(d)$, one can view the z as second-stage variables which can be adequately modified when the necessary information about load is available; the cost of tuning these is represented by $P(x_2(d))$, whose expectation is minimized along with the other objective function terms.

On the other hand, the risk term $E_D[P(d - \sum_i z_i)]$ can be viewed as a (continuous) averaged penalty function resulting from all possible deviations in uncertain demand.

In a general robust model (according to [20], [24]) the objective function would be composed of three terms:

$$E_\Omega[\xi] + \lambda Var_\Omega[\xi] + vE_\Omega[\sum Z_\omega^2],$$

where the first and second terms are expectation and variance operators for the total cost ξ with respect to a probability distribution of a given scenario set Ω. λ and v are the relative weights of the objective function terms, and Z_ω is the control variable of the constraints subject to uncertainty.

Our approach has certain similarities to robust optimization models, but an explicitly cost-variance term has been dropped from our objective function. In its place, we have included a new term that explicitly includes all the information about the probability distribution of the demand.

The penalty of infeasibilities to achieve solution robustness is more general in ECEM, in the sense that it allows us to consider (a) both discrete and continuous representation of the demand, and (b) the use of more general penalty functions P. In this sense, we consider it of interest to minimize the operating costs and the expected extra cost due to uncertainty, considering the probability distribution of the demand. Also, the outcome may be easier to assess here than it is for the solution provided by minimizing the scenario cost variance, which depends on the weight the user gives to the variance term.

Both ECEM and robust optimization models deal with uncertainty via formulations that provide robust strategies: In ECEM, robustness means that risks and costs are balanced, while in the models which include a variance term, robustness is a way of guaranteeing that the cost provided by any scenario does not differ too much from the average, since otherwise it is strongly penalized.

Note also that the resultant risk term in this work yields to a variety of penalties for infeasibility, not only the quadratic one. As an example, if we consider a uniform distribution of the demand and a linear penalty $P = k(d - \sum_i z_i)$, for a fixed $k > 0$, the resultant term is quadratic. If the demand were triangular, the risk term would be cubic, and so on for other penalty functions P and probability distributions.

One of the goals of this work is to enable the use of a wide range of probability distributions. We will show that the associated complexity in the risk term does not diminish the efficiency of the methodology, which is based on the explicit solution of the optimality conditions. Practical implications from a utility perspective include the broader range of possibilities for treating uncertainty and the large instances that may be solved. Another advantage of our approach is that the Benders-based methodology allows us to extend the model's scope to other stochastic-programming applications.

In the real operation of a electric power system the unit commitment is not determined until a very accurate demand forecast is available (hours or days ahead the real demand becomes known). The final generation and dispatching is a real time operation. In ECEM, each z_{its} is a generation level that provides the *a priori* least risk and serves to determine the optimum first-stage decisions x during the planning horizon, as well as the total expected investment and operating costs.

The "penalty function" employed by ECEM $P_{ts}\ (d_{ts} - \sum_i z_{its})$ may have different formulations for the different subperiod types, but it always depends on the difference between the observed value of the demand d_{ts} and the planned generation, $\sum_i z_{its}$. For low-demand subperiods, the formulation of P could be a simple linear function $P_{ts}(x_2) = kx_2$, for $x_2 = d_{ts} - \sum_i z_{its} > 0$. For example, this coefficient k could represent the cost per MWh of a coal thermal plant (because a low load level suggests that the generation outputs for thermal plants are below their generation capacities, so that the marginal costs will be low).

The penalty function and its interpretation, however, is different for subperiods corresponding to medium or peak load levels. Several interpretations can be derived from the risk term $E_D[P(x_2)]$ [i.e., the average value of the cost $P(x_2)$ due to deviations from the planned generation]. This means neither strict assessment of a feasible generation plan nor that the remaining demand (if any) is unmet. On the contrary, the cost $P(x_2)$ of adjusting the load curve could represent, among others: (a) the generation costs of existing fast start-up plants such as gas turbines (which have low capital costs but high operating costs), combined cycle or hydro plants (if the water values are high for those time periods), (b) energy purchase cost at the (open) spot market price, (c) need of additional fuel procurement at more costly prices, (d) energy purchase from other companies or interconnected countries, and (e) if unserved energy remains, the associated penalty.

The development of a more specific formulation concerns both the demand representation and the impact of the above features on cost, which depend on the specific electric system under consideration.

One issue of major importance is the way of deriving individual probability distributions for each subperiod of interest. We realize that some relationships may arise among related power demands (e.g., the demand in the peak subperiod of winter 1999 is not independent from the demand in the peak subperiod of winter 2000). The ideal case is when we can estimate the joint distribution of each random demand vector by parametric techniques (e.g., assuming an N-dimensional normal and estimating the mean vector and covariance matrix). Once this task has been accomplished, the marginal distributions related to each individual component (i.e., subperiod) are considered in defining the associated risk function. Model data may be preprocessed with a reasonable effort within the context of the problem considered.

3. GENERALIZED BENDERS DECOMPOSITION

ECEM is a large model for realistically sized problems, and requires decomposition to be solved. In this section, we present a Generalized Benders' Decomposition (GBD) scheme, which exploits the structure of the subproblem to accelerate convergence.

3. 1. Operation Subproblem

Suppose that $x = \hat{x}$ has been fixed. The remaining subproblem is usually known as the "operation subproblem." Removing the fixed cost $Q^x \cdot \hat{x}$ from the objective function, this subproblem can be formulated as

$$SP(\hat{x}): \qquad \min_{z \in \mathcal{Z}} Q^z \cdot z + R(z)$$

$$\text{s.t.} \qquad Az \leq -B\hat{x} + b^0(\mu),$$

where μ is the dual variable corresponding to available capacity.

In order to implement a Benders' decomposition scheme, solving the dual of $SP(\hat{x})$ is necessary. We take advantage of the lower triangular structure of A, and form one problem for each period t and subperiod $s \in S^t$, say,

$$SP_{ts}(\hat{x}): \qquad \min_{z_{1ts},\dots,z_{Its}} \sum_i Q^z_{its} z_{its} + W_{ts}\left(\sum_i z_{its}\right)$$

$$\text{s.t.} \begin{cases} z_{its} \leq b_{its}, & \forall i \ (\mu_{its}) \\ \sum_{i=1} z_{its} \leq m_{ts}, & \\ \sum_{i=1} z_{its} \geq l_{ts}, & \\ z_{its} \geq 0, & \forall i \ (u_{its}) \end{cases}$$

Note that if $\alpha_i = \beta_i$, the value of $b_{its} = a_{its} L_{ts} \tilde{x}_{it}$ depends only $x_{i1}, \ldots, x_{it}$ [see Eq. (2)]. In this case, the matrix A becomes the identity and $SP(\hat{x})$ can be separated into independent subproblems $SP_{ts}(\hat{x})$.

However, if $\alpha_i < \beta_i$, the expression above for b_{its} depends on the values of the z-variables for time periods $1, \ldots, t-1$. This fact makes $SP(\hat{x})$ nonseparable; we can adopt several solution strategies in this case. The first one would be to solve the $SP(\hat{x})$ as a nonlinear problem using a general-purpose methodology, but this could be difficult because of the large size of the subproblem for real cases. The second one consists of transforming the problem into a new one that enables to solve each $SP_{ts}(\hat{x})$ separately (by using the methodology described in the next section). This simplification can be done in two different ways. Since subproblem variables related to consecutive time periods are coupled, the first simplification consists of executing a relaxed-sequential algorithm which solves the t-time period subproblems by incorporating the solution provided by the previous periods $1, \ldots, t-1$. This iterative approach may not provide the optimal solution to the subproblem, but it is hoped that the accuracy of the dual variables will introduce violated Benders' cuts until a certain level of accuracy is achieved. In practice, Benders' cuts provided by a pseudooptimal solution of the subproblem work well, except for the last few iterations. The second strategy consists of redefining α_i as $\alpha_i = \beta_i$, so that the subproblem separability is preserved. With this assumption, we always consider the deterioration over time, whether the plant is being used or not. Obviously, other average rates of deterioration can be employed.

3.2. Solving $SP_{ts}(\hat{x})$

The $SP_{ts}(\hat{x})$ (and its dual) can be solved by means of the Kuhn–Tucker optimality conditions (KTOCs) associated with the subproblem. Detailed proof of all the results regarding this subproblem can be found in Salmerón [32]. Actually, the optimal solution is similar to the well-known merit order loading rule. The main difference is that, because of the penalty for failing to meet the demand, the plants should be dispatched in order of marginal running costs until this strategy exceeds the marginal penalty for failing to meet the load.

The solution [for a given $SP_{ts}(\hat{x})$] is obtained as follows:

- Assume $Q^z_{1ts} \le \cdots \le Q^z_{its} \le \cdots \le Q^z_{Its}$.
- Let $G(\tau) = dW(\tau)/d\tau$ be (the negative of) the marginal penalty function, and let $B_{rts} = \sum_{i=1}^{r} b_{its}$ be the cumulative available capacity of plants $i = 1, \ldots, r$.
- The optimal solution for the primal variables is

$$z_{its} = \begin{cases} b_{its}, & \text{for } i \le r, \\ k, & \text{for } i = r+1, \\ 0, & \text{for } i > r, \end{cases}$$

where r is the plant where the marginal operating cost exceeds the marginal penalty; i.e., r satisfies:

$$Q^z_{rts} \le -G(B_{rts}) \quad \text{but} \quad Q^z_{r+1,ts} > -G(B_{r+1,ts}).$$

The load supplied by the marginal plant is $k = G^{-1}(-Q^z_{r+1,ts}) - B_{rts}$, which is exactly the capacity level for plant $r+1$ whose penalty equals the cost of increasing the power output.

The dual variables take the following optimal values:

$$\mu_{its} = \left\{ \begin{array}{ll} Q^z_{r+1,ts} - Q^z_{its}, & \text{for } i \leq r \\ 0, & \text{for } i > r \end{array} \right\} \quad \text{and} \quad u_{its} = \left\{ \begin{array}{ll} 0, & \text{for } i \leq r \\ Q^z_{its} - Q^z_{r+1,ts}, & \text{for } i > r \end{array} \right\}.$$

The dual variables have a physical interpretation, as mentioned in [7]. For instance, $\mu_{its} > 0$ for $i \leq r$ represents the cost of reducing one unit of the available capacity in plant i, to be replaced by one unit from the marginal plant $r + 1$ (the last dispatched). On the other hand, $u_{its} > 0$ for $i > r$ indicates how the nondispatched plant i would increase system costs if it had to be dispatched.

REMARK 1: If the index r does not exist or the proposed solution does not satisfy $\sum_{i=1}^{r+1} z_{its} \geq l_{ts}$, it is necessary to use the merit order rule until the first plant r' satisfying $0 \leq z_{r'ts} = l_{ts} - B_{r'-1,ts} \leq b_{r'ts}$.

REMARK 2: Other special degenerate cases exist (e.g., when the marginal plant $r+1$ dispatches also its maximum capacity); see [32] for details.

REMARK 3: In Section 2.2, we mentioned that the operating cost function for a given plant i could be represented by a number of segments. This leads to a piecewise generation cost function at each plant i, in terms of the increasing heat rate costs:

$$Q^z_{its,1} \leq \cdots \leq Q^z_{its,K_i},$$

instead of a unique Q^z_{its}. However, this does not modify the way we obtain the solution of the subproblem for period (t, s), which in this case becomes

$$SP_{ts}(\hat{x}): \qquad \min_{z_{1ts,1},\ldots,z_{Its,K_I}} \sum_i \sum_{k=1}^{K_i} Q^z_{its,k} z_{its,k} + W_{ts}\left(\sum_i z_{its}\right)$$

$$\text{s.t.} \left\{ \begin{array}{ll} z_{its} \leq b_{its}, & \forall i \ (\mu_{its}), \\ z_{its} - \sum_{k=1}^{K_i} z_{its,k} = 0, & \forall i, \\ \sum_i z_{its} \leq m_{ts}, & \\ \sum_i z_{its} \geq l_{ts}, & \\ 0 \leq z_{its,k} \leq \bar{z}_{its,k}, & \forall i, k. \end{array} \right.$$

Note that the first, second and last constraints can be combined into a new one:

$$0 \leq z_{its,k} \leq b'_{its,k}, \qquad \forall i \ (\mu_{its,k}),$$

where

$$b'_{its,k} = \left\{ \begin{array}{ll} 0, & \text{if } \sum_{k'=1}^{k} \bar{z}_{its,k'} > b_{its}, \\ \min\left\{ \bar{z}_{its,k}, b_{its} - \sum_{k'=1}^{k-1} \bar{z}_{its,k'} \right\}, & \text{if } \sum_{k'=1}^{k} \bar{z}_{its,k'} \leq b_{its}. \end{array} \right.$$

Thus the total capacity b_{its} at plant i is divided into the different efficiency ranges k for the plant.

Finally, for every pair (t, s), we can sort the coefficients $Q_{its,k}$ to obtain the solution to $SP_{ts}(\hat{x})$ and its dual. Moreover, since for a fixed plant i the heat rate costs are ordered, it is guaranteed that $z_{its,k} > 0 \Rightarrow z_{its,k'} = \bar{z}_{its,k'}, \forall k' < k$, which preserves the consistency of the solution $z_{its} = \sum_{k=1}^{K_i} z_{its,k}$.

3. 3. Master Problem

ECEM can be represented by the following equivalent model with an infinite number of constraints:

$$\min_{x \in \mathcal{X}, x_0 \in \mathcal{R}} x_0$$

$$\text{s.t.} \begin{cases} x_0 \geq \min_{z \in Z} Q^x x + Q^z z + R(z) + \mu(Bx + Az - b^0), & \forall \mu \geq 0, \\ \min_{z \in Z} \lambda(Bx + Az - b^0) \leq 0, & \forall \lambda \in \Lambda, \end{cases} \tag{11}$$

where $\mu \in \Re^{I \cdot T \cdot S}$ and $\Lambda = \{\lambda \in \Re^{I \cdot T \cdot S}$ such that $\lambda \geq 0, \sum_{t=1}^{T} \sum_{s=1}^{S^t} \sum_i \lambda_{its} = 1\}$.

To solve this problem, the standard strategy is to relax the problem to a finite set of selected μ and $\lambda, \mu^1, \ldots, \mu^K$, and $\lambda^1, \ldots, \lambda^L$, where K and L are the cuts considered at each iteration. This model is usually referred to as the "Master Problem":

$$MP: \qquad \min_{x \in \mathcal{X}, x_0 \in \mathcal{R}} x_0$$

$$\text{s.t.} \begin{cases} x_0 - L^*(\mu^k, x) \geq 0, & \forall k = 1, \ldots, K, \\ L_*(\lambda^l, x) \leq 0, & \forall l = 1, \ldots, L, \end{cases}$$

where $L^*(\mu^k, x) = (Q^x + \mu^k B)x + \min_{z \in Z}\{(Q^z + \mu^k A)z + R(z)\} - \mu^k b^0$ and $L_*(\lambda^l, x) = \lambda^l Bx + \min_{z \in Z}\{\lambda^l Az\} - \lambda^l b^0$.

The first constraints $k = 1, \ldots, K$ are the so-called "type-1 cuts" or Benders optimality cuts, and the next $l = 1, \ldots, L$ are the "type-2 cuts" or Benders feasibility cuts.

Valid optimality cuts are provided by the dual solution of $SP(\hat{x})$, if feasible. If $K-1$ optimality cuts have been introduced, the new optimality cut (violated by the previous MP solution) has the form:

$$x_0 \geq (Q^x + \mu^K B)x + (Q^z + \mu^K A)\hat{z}^K + R(\hat{z}),$$

where $\hat{z}^K$ and μ^K are the optimal primal and dual solutions to $SP(\hat{x})$ at iteration K, respectively. In the case that $\hat{z}^K$ is not proved optimal due to $\alpha_i \neq \beta_i$ for some i (as discussed in Section 3.1), then the new $\hat{z}^K$ is defined as the optimum of

$$\min_{z \in Z}\{(Q^z + \mu^k A)z + R(z)\}$$

The solution to this model can be easily obtained, given that this is a separable problem that falls into the category of $SP(x)$ (actually it is a simpler model since the constraints associated with complicated variables are not considered). However, we cannot guarantee that following this procedure provides an active type-1 cut in the master problem.

Now we deal with the case where $SP(\hat{x})$ is infeasible. We check period $\tilde{t}$ and subperiod $\tilde{s}$ such that $SP_{\tilde{t}\tilde{s}}$ is infeasible. Since the only condition that may be violated is $\sum_i z_{i\tilde{t}\tilde{s}} \geq l_{\tilde{t}\tilde{s}}$, it can be proved that, by imposing $\sum_i b_{i\tilde{t}\tilde{s}} \geq l_{\tilde{t}\tilde{s}}$, we introduce a violated feasibility cut into the MP (or at least, a pseudofeasibility cut if $\alpha_i \neq \beta_i$ and $SP(\hat{x})$ is not solved exactly). This cut has the form

$$\sum_i a_{i\tilde{t}\tilde{s}} L_{\tilde{t}\tilde{s}}(\tilde{x}_{i\tilde{t}\tilde{s}} + b_i^0) \geq l_{\tilde{t}\tilde{s}}, \tag{12}$$

because feasibility in $SP_{\tilde{t}\tilde{s}}$ is achieved if and only if $\sum_i b_{i\tilde{t}\tilde{s}} \geq l_{\tilde{t}\tilde{s}}$. It can be shown (see [32]) that if $\alpha_i = \beta_i$, then (12) is an actual type-2 cut for the MP, and it is associated with the following vector λ in (11):

$$\lambda_{its} = \left\{ \begin{array}{ll} 1/I, & \text{if } t = \tilde{t} \text{ and } s = \tilde{s} \\ 0, & \text{otherwise} \end{array} \right\}$$

3. 4. Considerations for Solving the Master Problem

The MP is a linear problem with $I \cdot T + 1$ variables, x_0 and $x = (x_{11}, \ldots, x_{IT})$. The constraints $x \in \mathcal{X}$ and the cuts generated are the constraints included in the MP, which is solved via the dual simplex algorithm (Bazaraa and Jarvis [3]).

The main difficulty is to handle this problem when the number of cuts increases. A first simplification consists of selecting a few optimality cuts to be removed from the model (while still preserving the algorithm convergence) in order to control the size of the MP.

We have tested two possible ways for removing non-active constraints to create a ‘‘simplified MP’’ once MP^p (MP at iteration p) has been solved. The first way is to remove all non-active optimality constraints before solving MP^{p+1}; we call this ‘‘Total Elimination.’’ The second way is to remove only one nonactive constraint after solving MP^p (if such a constraint exists); we call this ‘‘Partial Elimination.’’ Note that in Partial Elimination, since at most one optimality cut is deleted, the number of optimality cuts is nondecreasing. In practice, the improvement in performance (total time until optimality) by using either procedure is more than 50% (even for small size problems) compared with keeping all the cuts.

There exist cases for which it is possible to improve the run time of the algorithm by taking into account knowledge about the formulation of feasibility cuts. In case that $\alpha_i = \beta_i \forall i$, it is clear (see Section 3.3) that at most $T \cdot S$ feasibility cuts defined in (12) will be necessary. Since the formulation of these cuts depends only on β_i, a_{its}, and l_{ts}, the MP can be started directly with some (or all) chosen cuts (i.e., for some or all periods and subperiods).

The advantage of introducing all possible feasibility cuts from the start is that this guarantees feasibility for all subsequent Benders’ subproblems. Unfortunately, this method only improves the performance for small problems. On the other hand, if $T \cdot S$ is large, few of these cuts will ever be active. In this case, it would be better to add only the necessary cuts by finding infeasible $SP_{ts}(\hat{x})$. An intermediate strategy is introducing a reduced subset of these cuts in advance. The choice may be, for example, to choose for each period $\tilde{t}$ the cut associated with the subperiod $\tilde{s}$ with the highest expected demand:

$$\tilde{s} = \text{argmax}\{l_{\tilde{t}s} : s = 1, \ldots, S\}.$$

Usually, it will be necessary to include other feasibility cuts while running the algorithm, but at least it is very likely that most of the selected cuts will be binding at the optimal solution.

4. COMPUTATIONAL RESULTS

In this section we present computational experience on the use of GBD for ECEM and the algorithms we have developed to solve the resultant submodels. We have solved ECEM using a Pentium 100 MHz computer with 8 Mb of RAM; the algorithm was coded in Fortran 77. Besides the Benders’ framework, we have also solved the model for smaller problems using the MINOS code as a benchmarking tool.

The problems that we have tested have been randomly generated according to the following characteristics:

- The number of plants I, periods T, and subperiods per period S are fixed in advance for each case.
- The distribution for random demands in each subperiod D_{ts} have been randomly chosen from the following list: (a) uniform, (b) triangular, and (c) normal. In those cases where the problem size allows us to compare with MINOS, only (a) or (b) has been selected.
- The penalty function $P_{ts}(\tau_{ts}, d_{ts})$ considered for each (t, s) is

  $$P_{ts}(\tau_{ts}, d_{ts}) = k_{ts}(d_{ts} - \tau_{ts}), \qquad \forall s = 1, \dots, S^t, \forall t = 1, \dots, T,$$

  where k_{ts} is also a positive and random real number.

  This function leads to function $G(\tau)$ (see Section 3.2) whose inverse must be calculated to solve the associated subproblems. In this case (dropping subscripts t and s), since

  $$W(\tau) = k\left(E_D[D] - \tau\int_\tau^\infty f(v)\,dv - \int_0^\tau vf(v)\,dv\right),$$

  we can obtain the partial derivatives of W as

  $$\frac{\partial W(\tau)}{\partial z_i} = G(\tau) = -k\int_\tau^\infty f(v)\,dv = -k\Pr\{D > \tau\}, \qquad \forall i = 1, \dots, I.$$

  This function is well known for a variety of probability distributions, so we can easily obtain G and its inverse at a given value of τ.
- The remaining costs and parameters of the matrix and rhs are also random data, following certain consistency rules according to the demand, a number of candidate plants, planning time, etc.
- For the availability coefficient a_{its}, we have noted that it has a strong influence on the algorithm performance. If it is ignored (i.e., assuming $a_{its} = 1$ for all i, t, s) the problem is much easier to solve than if $a_{its} < 1$. Some problems will have all the coefficients equal to 1. For others, we will assign a_{its} according to a uniform distribution on the interval (0.6, 1).

Table 1. Problem dimensions.

Problem	I	T	S per period	Linear vars.	Nonlin. vars.	Constr.	$a_{its} < 1$?
D5	5	5	5	25	125	180	Yes
D10	5	10	10	50	500	710	Yes
D11	5	11	8	55	440	627	No
D12	8	12	10	96	960	1212	No
D20	9	20	13	180	2340	2880	Yes
D32	9	32	20	288	5760	7072	Yes

Table 2. Problems with $\alpha = \beta$, CPU time for GBD method.

	GBD				
Problem	$E < 5\%$	$E < 1\%$	$E < 0.1\%$	$E < 0.01\%$	Best (B)
D5	$3''$	$5''$	$9''$	$15''$	116,350
D10	$13''$	$24''$	$38''$	$1'5''$	402,568
D11	$3''$	$7''$	$12''$	$15''$	335,781
D12	$15''$	$22''$	$51''$	$1'15''$	465,726
D20	$1'16''$	$2'52''$	$9'36''$	$17'21''$	945,389
D32	$6'21''$	$15'57''$	$1h4'$	$2h10'$	2,128,190

A brief description of the tested problems may be found in Table 1. Tables 2 and 3 concern problems where $\alpha_i = \beta_i$ and the algorithm's convergence is guaranteed. Table 2 shows the computational time using a GBD method to achieve a solution with a maximum error equal to $p\%$. In this table, the error is defined as $E = [(UB - LB)/LB] \cdot 100 < p$ (where UB and LB are the upper and lower bounds provided at each iteration). The algorithm is stopped when $E < p = 0.01\%$, and the best near-optimal solution is referred to as "Best" (B). This solution permits us to assess the actual time to achieve an error equal to $p\%$ (with respect to Best, not the gap provided dynamically by the Benders' algorithm); see Table 3. Also, we use Best to compare GBD with MINOS.

Now, let us consider the same problems but including some noise on the deterioration coefficients, so that $\alpha_i < \beta_i$ for a few plant types. We implement the relaxed-sequential algorithm to solve subproblems $SP(\hat{x})$ by incorporating the solution of previous time periods into the rhs of subsequent subproblems. The results (see Table 4) show that convergence is not guaranteed, and, for the first small problems, the solution provided by MINOS is better than the one provided by GBD without optimally solved subproblems. The necessary time to obtain a solution with $p\%$ error is evaluated according to the best known solution (B or M). Note that the matrix A is more dense here than in the other cases since $\alpha_i \neq \beta_i$.

Figures 1–3 plot some of the results obtained with GBD for problem $D20$. Figure 1 depicts the convergence of the lower and upper bounds in the Benders' algorithm.

In Figure 2, the number of optimality and feasibility cuts of the MP for each iteration is represented. We can see that, for the first iterations, only feasibility cuts are included (except the first optimality cut provided by the initial feasible solution). After some iterations, optimality and feasibility cuts are included. Eventually, only optimality cuts are necessary. For case $D20$ the necessary number of feasibility cuts at the optimal solution was 56. Note that before running the case we knew the number and formulation of all possible feasibility cuts (in total $20 \cdot 13 = 260$).

Table 3. Problems with $\alpha = \beta$.[a]

	GBD				MINOS			
Prob	$E < 5\%$	$E < 1\%$	$E < 0.1\%$	$E < 0.01\%$	$E < 5\%$	$E < 1\%$	$E < 0.1\%$	Best (M)
D5	$2''$	$4''$	$6''$	$11''$	$3''$	$4''$	$5''$	116,353
D10	$9''$	$20''$	$27''$	$51''$	$25''$	$30''$	$40''$	402,580
D11	$2''$	$6''$	$10''$	$14''$	$11''$	$17''$	$20''$	336,108
D12	$7''$	$18''$	$47''$	$1'15''$				No soln.
D20	$58''$	$1'45''$	$6'15''$	$12'51''$				No soln.
D32	$4'21''$	$10'54''$	$33'52''$	$1h10'$				No soln.

[a] Comparing GBD with MINOS given the Best (B) solution achieved. Best (M) is the best solution obtained with MINOS.

Table 4. Problems with $\alpha \neq \beta$.[a]

	GBD				MINOS			
Prob	$E < 10\%$	$E < 5\%$	$E < 1\%$	Best (B)	(B)-Gap	$E < 10\%$	$E < 5\%$	Best (M)
D5	$4''$	$5''$	$7''$	116,421	2.49	$3''$	$3''$	115,221
D10	$1'15''$	$1'20''$	$2'0''$	385,217	1.49	No	No	433,499[b]
D11	$10''$	$15''$	$25''$	332,736	0.15	$30''$	No	349,536[b]
D12	$20''$	$35''$	$1'10''$	459,482	0.27			No soln
D20	$9'55''$	$11'27''$	$18'50''$	901,370	0.74			No soln
D32	$6'45''$	$13'0''$	$31'15''$	2,010,192	0.41			No soln

[a] Comparing GBD with MINOS for the best solution found (B or M). The (B)-Gap is provided by the Benders' algorithm (%).

[b] MINOS provided a feasible solution but crashed before ending the optimization.

This example shows that it would not be a good strategy to include all these constraints from the beginning since most of them are unnecessary. However, we included 20 out of them (one per period, corresponding to the subperiod with highest expected demand) to help the algorithm in finding the necessary type-2 cuts. Regarding optimality cuts, they are updated in a list, so that when a new optimality cut must be considered, it replaces the first nonactive optimality cut of the previous iteration (if it exists; otherwise the list size is increased). This allow us to perform a large number of iterations and control the dimension of the MP.

Figure 3 depicts the cpu time to run MP, SP, and the whole algorithm (including other features such as data processing). The first iterations are not time-consuming since $SP(\hat{x})$ is infeasible. When $SP(\hat{x})$ becomes feasible, the necessary time to solve it is similar for all iterations because the number of subproblems $SP_{ts}(\hat{x})$, and computational operations to be performed, is always the same. The time spent on the solving the MP is close to a linear function because of the cut-elimination scheme used.

Finally, Figure 4 shows a comparison of convergence rates between MINOS and GBD for Problem D11.

5. CONCLUSIONS

We have formulated a stochastic capacity-expansion model. It takes into account particular features of electric systems such as deterioration and availability over time. Deterioration affects

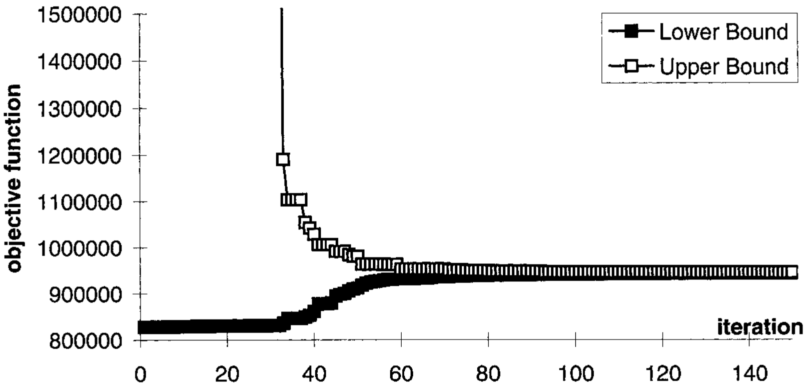

Figure 1. Benders' performance for problem $D20$.

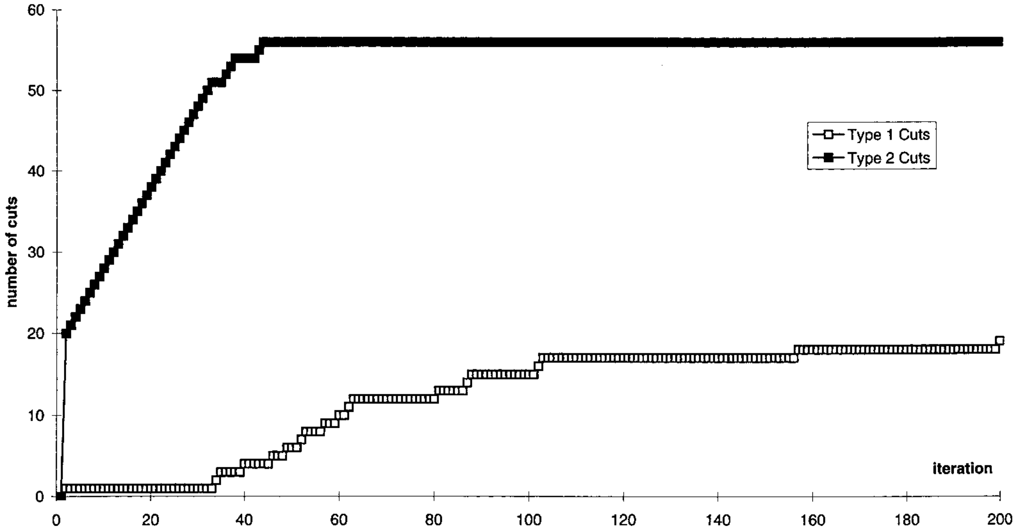

Figure 2. Benders' cuts for problem $D20$.

the generation units (turbines or thermal groups), having a progressive effect on the generation capacity. Availability is a random variable that depends on external factors (such as water inflow, sunlight hours, shutdown needs for maintenance, and so on) which can be estimated with historical data.

The most important aspect considered in this work is the treatment of the demand uncertainty. We deal with this issue by creating a risk-based function as part of the model's objective function.

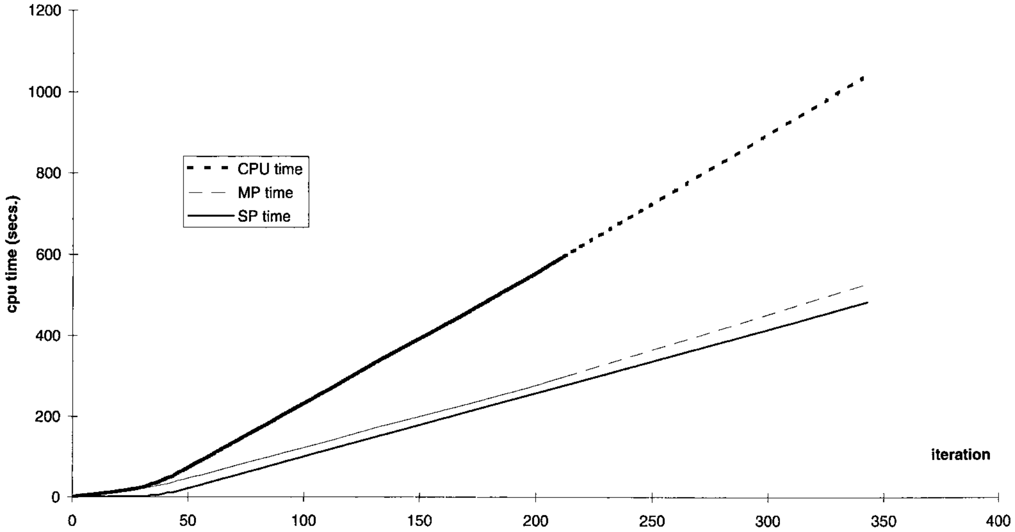

Figure 3. CPU time for problem $D20$.

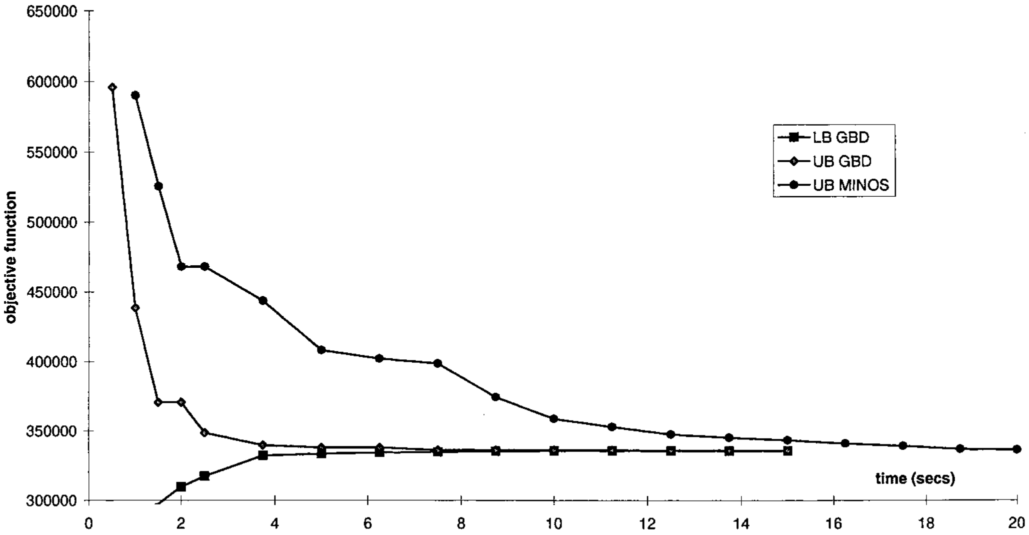

Figure 4. GBD and MINOS for problem $D11$.

We have discussed our model's goal and potential value, and have compared the formulation to others in the existing literature. We focused part of the discussion on pointing out the benefits and drawbacks in relation to standard robust optimization models.

A Benders' decomposition algorithm to solve the model has been implemented. The Benders' subproblem is highly nonlinear, but we can exploit its structure and solve it explicitly. The necessary calculations depend linearly on the number of plants, periods and subperiods. This fact improves the algorithm performance compared with other methods.

Several small cases have permitted comparison with the use of standard packages like MINOS. In our computational experience the proposed techniques have yielded good results which suggest retaining this scheme in future extensions of the work.

The methodology can be extended to other capacity-expansion models incorporating other sources of uncertainty. A more realistic approach considering a discrete set of possible values for the capacity to be installed is part of ongoing research. The idea is to preserve the advantages of the subproblem structure, although the master problem becomes a more complicated mixed-integer model. Since the addressed Benders' subproblem has a structure that appears in other models, the proposed methodology may also be used in other applications of interest.

ACKNOWLEDGMENTS

The authors wish to thank three anonymous referees, the associate editor, and Professor Kevin Wood of Naval Postgraduate School for their suggestions and helpful comments on this paper. J. Salmerón wishes to thank the Naval Postgraduate School for the support provided during the last part of this work.

REFERENCES

[1] K.A. Ariyawansa and D.D. Hudson, Performance of a benchmark parallel implementation of the Van Slyke and Wets algorithm for two-stage stochastic programs on the Sequent/Balance, Concurrency Practice Experience 3 (1991), 109–128.

[2] M. Álvarez, C. Cuevas, L. Escudero, J. De la Fuente, C. García, and F. Prieto, Network planning under uncertainty with an application to hydropower generation, Top 2 (1994), 25–58.

[3] M. Bazaraa and J. Jarvis, Linear programming and network flows, Wiley, New York, 1977.

[4] O. Berman, Z. Ganz, and J. Wagner, A stochastic optimization model for planning capacity expansion in a service industry under uncertain demand, Nav Res Logistics 41 (1994), 545–564.

[5] J. Birge and D. Holmes, Efficient solution of two-stage stochastic linear programming using interior point methods, Comput Optim Appl 1 (1992), 245–276.

[6] J. Birge and F. Louveaux, Introduction to stochastic programming, Springer-Verlag, New York, 1997.

[7] J.A. Bloom, Solving an electricity generation expansion planning problem by generalized Benders' decomposition, Oper Res 31 (1983), 84–100.

[8] J.A. Bloom, M. Caramanis, and L. Charny, Long-range generation planning using generalized Benders' decomposition: Implementation and experience, Oper Res 32 (1984), 290–313.

[9] W. Dapkus and T. Bowe, Planning for new electric generation technologies. A stochastic dynamic programming approach, IEEE Trans Power Appl Syst PAS-103 (1984), 1447–1453.

[10] M. De Groot, Optimal statistical decisions, McGraw-Hill, New York, 1970.

[11] Y. Ermoliev and R. Wets, Numerical techniques for stochastic optimization, Springer-Verlag, New York, 1988.

[12] L. Escudero, I. Paradinas, J. Salmerón, and M. Sánchez, SEGEM: A simulation approach for electric generation management, IEEE Trans Power Syst 13 (1998), 738–748.

[13] A. Geoffrion, Generalized Benders decomposition, J Optim Theory Appl 10 (1972), 237–260.

[14] B. Golun, M. Holmer, R. McKendall, L. Pohlman, and S. Zenios, A stochastic programming model for money management, Eur J Oper Res 85 (1995), 282–296.

[15] B. Gorenstein, N. Campodonico, J. Costa, and M. Pereira, Power system planning under uncertainty, IEEE Trans Power Syst 8 (1993), 129–136.

[16] R. Hiller and J. Eckstein, Stochastic dedication: Designing fixed income portfolios using massively parallel Benders' decomposition, Manage Sci 39 (1993), 1422–1442.

[17] B. Hobbs, Optimization methods for electric utility resource planning, Eur J Oper Res 83 (1995), 1–20.

[18] D.R. Luce and H. Raifa, Games and decisions, Wiley, New York, 1967.

[19] S. Malcolm and G. Anandalingam, Robust optimization for power systems capacity expansion planning in large countries, Working Paper 98-09, Department of Systems Engineering, University of Pennsylvania, Oper Res, in press.

[20] S. Malcolm and S. Zenios, Robust optimization for power systems expansion under uncertainty, J Oper Res 45 (1994), 1040–1049.

[21] A. Marín and J. Salmerón, Electric capacity expansion under uncertain demand: Decomposition approaches, IEEE Trans Power Syst 13 (1998), 333–339.

[22] J. Mulvey and H. Vladimirou, Stochastic network programming for financial planning problems, Manage Sci 38 (1992), 1642–1664.

[23] J. Mulvey and A. Ruszczynski, A new scenario decomposition method for large-scale stochastic optimization, Oper Res 43 (1995), 477–490.

[24] J. Mulvey, R. Vanderbrei, and S. Zenios, Robust optimization for large-scale systems, Oper Res 43 (1995), 264–281.

[25] F. Murphy and Z. Wang, A network reformulation of an electric utility expansion planning model, Nav Res Logistics 40 (1993), 451–457.

[26] F. Noonan and R. Giglio, Planning electric power generation: A nonlinear mixed integer model employing Benders decomposition, Manage Sci 23 (1977), 946–956.

[27] M. Pereria and L. Pinto, Multi-stage stochastic optimization applied to energy planning, Math Program 52 (1991), 359–375.

[28] S. Rajagopalan, Deterministic capacity expansion under deterioration, Manage Sci 38 (1992), 525–539.

[29] R. Rockafellar and R. Wets, Scenarios and policy aggregation in optimization under uncertainty, Math Oper Res 16 (1991), 119–147.

[30] A. Ruszczynski, Parallel decomposition of multistage stochastic programming problems, Math Program 58 (1993), 201–228.

[31] A. Ruszczynski, Decomposition methods in stochastic programming, Math Program 79 (1997), 333–353.

[32] J. Salmerón, Optimización en diseño de grandes sistemas. Aplicaciones eléctricas, Ph.D. thesis, Dept Mat Aplicada y Estad., Univ. Politécnica de Madrid, 1998.

[33] D. Sherali and K. Staschus, A two-phase decomposition approach for electric utility capacity expansion planning including nondispatchable technologies, Oper Res 38 (1990), 773–791.

[34] H. Sherali, K. Staschus, and J. Huacuz, An integer programming approach and implementation for an electric utility capacity planning problem with renewable energy sources, Manage Sci 33 (1987), 831–847.

[35] H. Vladimirou, Computational assessment of distributed decomposition methods for stochastic linear programs, Eur J Oper Res 108 (1998), 653–670.

[36] R. Wets, ‘‘Stochastic programming,’’ Handbooks in OR & MS, Vol. 1, G.L. Nemhauser et al., Eds., Elsevier, Amsterdam, 1989.

[37] W. Winston, Operations research: Applications and algorithms, PWS Publishers, 1987.

[38] D. Yang and S. Zenios, A scalable parallel interior point algorithm for stochastic linear programming and robust optimization, Comput Optim Appl 7 (1997), 143–158.