Stochastic optimization approach to water management in cooling-constrained power plants

Juan M. Salazar a, Urmila Diwekar a,⇑, Emil Constantinescu b, Victor M. Zavala b

a Vishwamitra Research Institute, Center for Uncertain systems: Tools for Optimization and Management, Clarendon Hills, IL 60439, USA
b Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA

HIGHLIGHTS

• Addresses the problem water shortages and effect weather conditions on plant performance.
• Presents the problem as a stochastic programming problem.
• Uses an efficient new algorithm BONUS to solve the problem.
• Provides real time optimization solutions for various scenarios.

ARTICLE INFO

Article history:
Received 11 November 2012
Received in revised form 16 April 2013
Accepted 28 May 2013

Keywords:
Stochastic optimization
Power plants
Weather forecasting
Uncertainty
Water constraints

ABSTRACT

A stochastic optimization framework for water management in cooling-constrained power plants is proposed. The approach determines optimal set-points to maximize power output in the presence of uncertain weather conditions and water intake constraints. Weather uncertainty is quantified in the form of ensembles using the state-of-the-art numerical weather prediction model WRF. The framework enables the handling of first-principles black-box simulation models by using the reweighting scheme implemented in the BONUS solver. In addition, it enables the construction of empirical distributions from limited samples obtained from WRF. Using these computational capabilities, the effects of cooling constraints and weather conditions on generation capacity are investigated. In a pulverized coal power plant study it has been found that weather fluctuations make the maximum plant output vary in the range of 5–10% of the nominal capacity in intraday operations. In addition, it has been found that stochastic optimization can lead to daily capacity gains of as much as 245 MW h over current practice and enables more robust bidding procedures. It is demonstrated that reweighting schemes can enable real-time implementations.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

The U.S. geological survey estimates that thermoelectric generation accounts for approximately 500,000 million liters per day of fresh water withdrawals, ranking only slightly behind agricultural irrigation as the largest source of freshwater withdrawals in the United States [1]. In base-load power plants (i.e., coal and nuclear) huge amounts of water are lost to the environment in the cooling towers as a result of evaporation. This water loss needs to be extracted by the process from nearby aquifers and/or publicly owned treatment facilities. Current technologies based on cooling tower systems utilize fresh water. Alternatively, one can use dry cooling and non-freshwater resources. However, such use results in prohibitive cost penalties to the power system [2,3]. A 500 MW coal-fired power plant that employs once-through cooling can use more than 50 million liters per hour of water for cooling and other process requirements. The latest forecast estimates of the Energy Information Administration indicate that U.S. coal-fired generating capacity will grow from approximately 305 GW in 2004 to 453 GW in 2030. This growth is a concern not only because of increasing emissions but also because of decreasing freshwater levels and the increasing frequency and severity of drought conditions [4,5].

Water consumption in power plants is intimately related to weather conditions such as ambient temperature and humidity because these conditions affect the capacity of the cooling towers. For instance, independent system operators normally represent plant capacity as a function of ambient temperature in clearing procedures. Unfavorable weather conditions can force power plants to drop their power outputs or even to fully shut down, thus limiting
their participation in electricity markets. Consequently, weather conditions can have a strong effect on market prices and grid reliability particularly if the plant is located in an area with significant transmission congestion [6]. Reduced water availability and increasing environmental regulation suggest that a more restricted water intake will be enforced in power plants in order to ensure long-term availability. This situation will force cooling systems to operate closer to their maximum capacity and will make them highly sensitive to weather conditions. Consequently, quantifying the uncertainty of the ambient conditions and modifying operations in real time will become increasingly important.

Stochastic optimization has been widely used to coordinate generation at central grid levels [7–10]. Studies at the power plant level have been focused mostly on determining optimal bidding strategies [11]. These studies, however, neglect detailed physical limitations of the power plant. Recently, a stochastic optimization approach has been proposed to minimize water consumption in power plants [12,13]. There, the authors determine optimal operating conditions (set-points for the control system) for a pulverized coal power plant that minimize water consumption under fluctuations of load and seasonal ambient conditions. Empirical distributions for ambient temperature and relative humidity were constructed by using historical weather data. The approach presented in that paper (and the current paper) is applicable beyond the pulverized coal power system. A similar approach can be used for water management in advanced power systems like integrated gasification combined cycle systems, renewable energy based power systems, and other water-constrained manufacturing plants.

In this work, the numerical weather prediction system WRF is used to obtain forecast and uncertainty information [14,10]. This enables the determination of physical representations of uncertainty and to investigate their size, shape, and accuracy. A specialized resolution targeting scheme is used in WRF to enable fine resolutions at the location of the power plant. This approach has been used in diverse studies to obtain weather information affecting energy system performance [15,16]. A detailed case study is presented that integrates uncertainty quantification with WRF, stochastic optimization, and a detailed power plant first-principles model to investigate computational feasibility, effects of weather, and uncertainty levels on plant performance and to evaluate potential over the state of the art. This is the first study that integrates physical weather and power plant models. The study demonstrates that stochastic optimization can yield savings in daily plant capacity of more than 245 MW h over current practice. In addition, this approach enables more robust bidding procedures.

The BONUS stochastic optimization solver is used. The solver implements a reweighting-based algorithm to solve the associated nonlinear stochastic optimization problem [17]. Reweighting is a powerful paradigm for handling black-box simulation models because it enables the construction of empirical distributions of the objective function off-line that can be updated on-line along the search. This avoids the need to perform simulations on-line; hence, it dramatically decreases computational times and ultimately enables feasible industrial implementations.

The paper is structured as follows. Section 2 presents a high-level description of power plant operations and effects of weather and water intake constraints are presented. Section 3 presents a stochastic optimization formulation and the proposed solution strategy. Section 4 presents the weather forecasting system WRF and the procedure used to compute physical uncertainty representations in the form of ensembles. The integrative case study is described in Section 5 and, finally, concluding remarks and a discussion of future work are provided in Section 6.

### 2. Power plant operations

This section presents a high-level description of cooling-constrained power plants. In addition, the effect of weather conditions and water constraints on generation capacity is explained.

Fig. 1 presents a typical interface between a typical power plant equipped with a wet recirculating cooling system. For a detailed description of power plant operations and other types of cooling systems the reader is referred to [18]. The power plant burns fuel to generate steam that is used to generate power in a turbine train. The steam coming out of the turbines must be condensed before being recirculated to the power plant. This steam has a thermal load at time \( t \) denoted by \( Q_g(t) \).

The thermal load depends on the total amount of power produced by the plant \( P(t) \), which implicitly sets the amount of steam to be condensed. The plant output variables of interest are denoted...
as $x_6(t)$. In this case, these are given by $P(t)$ and $Q_c(t)$. The operational degrees of freedom of the power plant are denoted as $u_6(t)$. These can be the set-points, fuel flow rate, stoichiometric ratios, and so on. It is assumed that the plant model does not exhibit coupling in time (i.e., it is at steady-state). Consequently, the explicit dependence in time from the nomenclature are removed. A plant model in the following abstract form is considered:

$$0 = f_p(x_6, u_6). \quad (2.1)$$

The cooling system capacity is given by $Q_c$. This is part of the cooling system states denoted by $x_c$. The cooling system capacity depends on the weather conditions (relative humidity and dry bulb temperature) denoted as $w$. The operational degrees of freedom of the cooling system are denoted as $u_c$. The cooling system model has the following form:

$$0 = f_c(x_c, u_c, w). \quad (2.2)$$

At steady-state, the cooling system capacity should match that required by the power plant. This condition is given by the abstract condenser model of the form

$$0 = g(x_p, x_c). \quad (2.3)$$

The capacity of the cooling system can be denoted in implicit form as $Q_c(u_c, w)$. The maximum capacity by design under ideal weather conditions are defined as $Q_c^{\text{max}}$. This maximum capacity is constrained by physical limitations of the system such as diameter, height, flooding conditions, and pumping capacity, by environmental regulations, and by drought. These limitations typically manifest as water constraints, which limit the amount of fresh water (make-up water) that the cooling system can use to provide cooling capacity. For any other nonideal weather conditions the condition $Q_c(u_c, w) \leq Q_c^{\text{max}}$ holds. The feasible operating region for the cooling system is given by the set of inequality constraints,

$$0 \leq h_i(x_c, u_c, w), \quad (2.4)$$

which depend on the weather conditions and design specifications. This set of constraints can also include bounds on states and decision variables for the cooling system. The feasible operating region for the power plant is denoted as an inequality system of the form

$$0 \leq h_j(x_p, u_p, w). \quad (2.5)$$

This set of constraints includes bounds on states and decision variables for the plant. The entire model coupling the power plant and the cooling system can be written as follows:

$$0 = f(x, u, w), \quad (2.6a)$$

$$0 \leq h(x, u, w). \quad (2.6b)$$

Water management is important in plant operations as it can actually dictate the maximum amount of power but this is not always the case. Assume that the maximum power output under design specifications (nominal capacity) is given by $P^{\text{max}}$, with corresponding cooling demand $Q_c^{\text{max}}$. If the available cooling capacity at current conditions satisfies $Q_c(u_c, w) \geq Q_c^{\text{max}}$, then the power output is constrained only by the plant-side design specifications (e.g., furnace capacity) and is given by $P^{\text{max}}$. If the cooling capacity is constrained $Q_c(u_c, w) < Q_c^{\text{max}}$, then $P^{\text{max}} = P^{\text{max}}(Q_c(u_c, w))$, and, in order to have feasible operation, it is required that $P \leq P^{\text{max}}(Q_c(u_c, w))$ and $Q \leq Q_c^{\text{max}}(Q_c(u_c, w))$. In other words, the maximum capacity of the power plant is an implicit function of the available cooling capacity. In this case, the power plant is operationally constrained by the prevailing weather conditions because the cooling capacity is, in turn, constrained by weather and regulations (e.g., maximum discharge temperature). Therefore, there exists a feasibility threshold value $Q_c^{\text{max}}$ where the plant moves from being design constrained to cooling constrained and water management is essential. This also implies that the actual sensitivity of the maximum power to cooling capacity is affected by the inherent design of the power plant. If the power output is constrained by furnace capacity, no sensitivity of the maximum power output to cooling capacity will be observed, and water management becomes less important. In most designs, however, cooling does represent an economic constraint because of the large amounts of fresh water required. In addition, it is key to adaptively identify the cooling constrain threshold in order to ensure future operational feasibility in the face of uncertain weather or regulation conditions.

In this study, it is assumed that the main operational objective in the power plant is to maximize plant output power at prevailing weather conditions. To achieve this objective, one can adjust the set-points represented by $u$. For known weather conditions, the optimal set-points can be determined by solving a nonlinear optimization problem of the form

$$\begin{align*}
\max_{u} & \quad J(x, u, w) \\
\text{s.t.} & \quad 0 = f(x, u, w) \\
& \quad 0 \leq h(x, u, w).
\end{align*} \quad (2.7a-2.7c)$$

Here, the function $J(x, u, w)$ can be a general objective function. In the proposed setting, however, it will simply denote the plant output power. The set-points on the plant side can be manipulated to minimize the cooling demand of the plant in order to avoid the combined effect of weather and water constraints. In other words, if the set-points are not adapted to the prevailing weather conditions, the power plant output will be constrained by the water intake constraints, and the power output will decrease.

The steady-state modeling assumption implies that the distance between time steps is sufficiently long such that there exists a low-level control system capable of making the dynamic transition. This assumption is justified for slow base-load plants with time constants on the order of hours.
3. Stochastic optimization

In this section a stochastic optimization formulation and solution strategy are presented for mitigating effects of weather on power plant capacity.

3.1. Formulation

The objective of the stochastic formulation is to find set-points for the current time $t$, denoted by $u_{HN}$, that maximize the expected value of the plant capacity. The weather conditions $w$ are assumed to be random variables.

$$\max_u \mathbb{E}_w[f(x(w), u, w)]$$

s.t. $0 = f(x(w), u, w)$, $w \in \Omega$

$$0 \leq h(x(w), u, w), w \in \Omega.$$  

(3.8a)

(3.8b)

(3.8c)

Here, $\mathbb{E}_w[\cdot]$ denotes the expected value of the objective function $f(\cdot)$ with respect to $w$. The set $\Omega$ is the support of the distribution of $w$. In the context of the presented work, this is the expected output power of the power plant. The constraints are assumed to hold almost surely (for every scenario). The optimal value of this problem (the expected value) is denoted as $V(u_{HN}) := \mathbb{E}_w[f(x(w), u_{HN}, w)]$, where the subscript $HN$ indicates the here-and-now nature of the solution $u_{HN}$. In other words, a single (scenario independent) set of implementable set-points is sought. In practical terms, this set-point ensures the feasibility of the model and constraints for each realization of the weather conditions $w \in \Omega$. Note that the state variables are scenario-dependent and thus represented as $x(w)$. Also note that since the model is assumed to be in steady-state, the problems at different time instants are decoupled. Consequently, the above problem can be solved over an horizon $t, \ldots, t + T$ to determine the expected values $V(u_{HN}(\tau)), \tau = t, \ldots, t + T$. Related to the here-and-now optimization problem (3.8) is the deterministic problem:

$$\max_u f(x, u, \mathbb{E}[w])$$

s.t. $0 = f(x, u, \mathbb{E}[w])$

$$0 \leq h(x, u, \mathbb{E}[w]).$$

(3.9a)

(3.9b)

(3.9c)

This problem uses the expected value of the weather conditions $\mathbb{E}[w]$ to determine the optimal set-points, which is denoted as $u_N$. If the set-points $u_N$ are fixed and the objective function for each scenario $w \in \Omega$ is evaluated, one can obtain $f(x(w), u_N, w)$ and compute the deterministic expected value

$$V(u_N) := \mathbb{E}_w[f(x(w), u_N, w)].$$

(3.10)

From [7] the relationship

$$V(u_{HN}) \geq V(u_N)$$

(3.11)

is obtained. Consequently, the stochastic optimization solution will always give better performance than will the deterministic counterpart (as long as the global maximum can be computed). The improvement (normally called the value of the stochastic solution), however, is problem dependent and thus can be determined only empirically. In the power plant context this value is determined by the design specifications and the ambient conditions at the particular climate and season. In this work, this value is determined for a realistic setting.

The effect of using weather forecasting to optimize operating conditions in real time (intraday) is also of interest in this study. To this end, the solution of the deterministic problem is considered using day-average weather conditions, represented by $w$. This gives a single fixed set-point for the entire time horizon $t, \ldots, t + T$ denoted by $u_D$ and associated expected values $V(u_D(\tau)), \tau = t, \ldots, t + T$. The objective function improvement with respect to the deterministic objective function $V(u_N)$ gives a measure of the impact of using weather forecast information to perform intraday optimization of operating conditions. This is important because, in a typical market setting, the policy of set-points over the entire horizon $t, \ldots, t + T$ must be determined in advance (normally within 24 h) in order to determine a plant capacity profile to be bid into the market. In other words, inaccurate estimation of weather conditions can ultimately affect plant profit as well as grid reliability and market prices because it can lead to under- and overestimation of capacity.

3.2. Solution

The stochastic optimization problem (3.8) can be written in the following abstract form:

$$\min_{u \in \mathcal{U}} V(u) := \mathbb{E}_w[\phi(u, w)].$$

(3.12)

Here, $\mathcal{U}$ is the box set $\mathcal{U} := \{u \leq u \leq \bar{u}\}$, where $u$ and $\bar{u}$ are lower and upper bounds, respectively. The function $\phi(u, w)$ results from evaluating $f(x(w), u, w)$ at a given $x(w)$, $w$ that solves the model $f(x(w), u, w) = 0$. In other words, this representation can be used if it is assumed that inequality constraints exist only in $u$ and if it is assumed that the implicit solution of the black-box model $f(x(u, w) = 0$ can be represented as a smooth function of $u$ and $w$, $x(u, w)$. This is the case considered here.

The random vector $w$ has a prior distribution $\mathcal{P}(w)$. Here, the term prior indicates that it is exogenous or independent to the plant. Note that the objective function $\phi(u, w)$ also has an associated posterior distribution $\mathcal{P}(\phi(u, w))$ from which its expected value will be maximized. The term posterior indicates its implicit functional dependency on the independent variables $u, x$. The stationarity (Karush–Kuhn–Tucker or KKT) conditions of this problem are

$$0 = \nabla_u V(u) + \psi - \lambda,$nabla_u V(u) + \psi - \lambda,$

(3.13a)

$$0 \leq \psi \perp (\bar{u} - u) \geq 0,$nabla_u V(u) + \psi - \lambda,$

(3.13b)

$$0 \leq \psi \perp (u - \bar{u}) \geq 0.$$nabla_u V(u) + \psi - \lambda,$

(3.13c)

Here, $\lambda$ and $\psi$ are multipliers for upper and lower bounds, respectively. To solve the stochastic optimization problem, one can follow exterior or interior sampling approaches. In exterior sampling approaches (also known as sample-average approximation approaches) $N_{sam}$ samples are drawn from the prior distribution $\mathcal{P}(w)$ to convert the problem (3.12) into a deterministic counterpart [19] of the form

$$\min_{u \in \mathcal{U}} \mathcal{V}(u) := \frac{1}{N_{sam}} \sum_{i=1}^{N_{sam}} \phi(u, w_i)).$$nabla_u V(u) + \psi - \lambda,$

(3.14)

Inference is performed a posteriori to check the quality of the solution because the objective function $\mathcal{V}(u)$ is an empirical approximation of $V(u)$ [20]. One of the issues with the exterior sampling approach is that the number of samples normally must be overestimated in order to guarantee an appropriate solution accuracy. This is an issue of concern because, in many applications, only a few samples are available and because the complexity of the deterministic counterpart increases with the number of samples. For instance, sampling of the prior distribution might require the solution of computationally intensive model (e.g., a weather forecasting model); or, if the prior is a high-dimensional normal, sampling will require the factorization of a large and dense covariance matrix (cubic complexity). In [10], the authors proposed a reweighting scheme to estimate additional samples from the limited number of samples obtained from a weather forecasting system.
Interior sampling approaches estimate the objective function and derivative information internally along the search. One of the key advantages is that flexibility can be used to minimize the number of samples needed by identifying promising solutions early in the search. This concept was implemented by Sahin and Diwekar in the BONUS solver [17]. The authors proposed a reweighting scheme to estimate directly the posterior distribution of the objective function \( \mathcal{P}(\phi(u, w)) \) and associated expected value and derivative information. This is different from the reweighting scheme presented in [10], which was applied to the prior distribution of the random variables.

Reweighting can also be interpreted as an importance sampling approach [21]. Using this observation, one can estimate the statistical properties of the objective function posterior indirectly by updating design distributions of the random and input variables and the objective function distribution generated off-line (before running the optimization search). This is particularly useful in a simulation-based approach where the evaluation of the objective function requires the simulation of a black-box model. The approach contrasts with traditional sampling algorithms that simulate the model along the search to estimate the expected objective function and gradient information. Consequently, the internal reweighting approach can save large amounts of on-line computational time. In addition, reweighting enables the use of highly complex models (e.g., using internal convergence loops) and black-box legacy codes. Further, the use of reweighting enables prior distributions to be created using samples from another distribution that is more expensive to sample. For these reasons, the internal reweighting approach implemented in the BONUS solver is used. A detailed explanation of the procedure can be found in [17]. This is essentially a sequential quadratic programming (SQP) that uses a reweighting scheme to estimate the objective function and gradient. The Hessian is approximated by using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula.

The procedure involves the following general steps:

1. **Off-Line Computations.** Draw independently distributed samples \( j = 1, \ldots, N_{\text{sam}} \) for random variables \( w_j \) and inputs \( u_j \) covering the convex set \( \mathcal{U} \). Use these samples to generate the design prior density \( \mathcal{P}_d(u, w) \) using kernel density estimation (KDE). Evaluate the objective function for each sample \( j = 1, \ldots, N_{\text{sam}} \) to obtain the objective function response \( \phi(u_j, w_j) \), and estimate the design posterior distribution \( \mathcal{P}(\phi(u, w)) \) using KDE.

2. **On-Line Computations.** Given the prior distribution for the random variables \( \mathcal{P}(w) \) and the design distributions \( \mathcal{P}_d(u, w), \mathcal{P}(\phi(u, w)) \) and objective values \( \phi(u_j, w_j), j = 1, \ldots, N_{\text{sam}} \), initialize decision variables \( u^k \).

   - **2.1. Optimality Check.** At iteration \( k \), and using the current iterate of decision variables \( u^k \), define a narrow normal distribution around this point, and draw samples \( u^k \) from it. Draw samples \( w^k \) from the available distribution. Use samples to generate \( \mathcal{P}_d(u, w) \) using KDE. Estimate the objective function using the reweighting formula:

   \[
   \bar{V}(u^k) \approx \mathbb{E}_w[\phi(u^k, w)] = \sum_{j=1}^{N_{\text{sam}}} \alpha_{j}^k \phi(u_j, w_j),
   \]

   where the weights \( \alpha_{j}^k \) are obtained from

   \[
   \alpha_{j}^k = \frac{\mathcal{P}_d(u_j, w_j) \mathcal{P}(\phi(u_j, w_j))}{\sum_{i=1}^{N_{\text{sam}}} \mathcal{P}_d(u_i, w_i) \mathcal{P}(\phi(u_i, w_i))}
   \]

   and satisfy \( \sum_{j=1}^{N_{\text{sam}}} \alpha_{j}^k = 1 \). Perturb the decision variables \( u^k + \Delta u^k \) to estimate the perturbed objective function \( \bar{V}(u^k + \Delta u^k) \) using reweighting to estimate the gradient \( \nabla_u \bar{V}(u^k) \). If the KKT residual \( \| \nabla_u \bar{V}(u) + v - \psi \| \) and complementarity products are sufficiently small, terminate. Otherwise, go to step 2.2.

   - **2.2. SQP Step Computation.** Use the gradient to compute the Hessian approximation \( H_k \) using the BFGS formula, and compute step \( \Delta u_k \) for decision variables by solving the quadratic program (QP):

   \[
   \min_{\Delta u} \nabla_u \bar{V}(u^k)^T \Delta u + \Delta u^T H_k \Delta u,
   \]

   s.t. \( u^k + \Delta u \in \mathcal{U} \).

   Cut the step if necessary to obtain a new iterate \( u^{k+1} = u^k + \alpha \Delta u \) with \( \alpha \in (0, 1] \). Go to step 2.1.

As can be seen, the computation of the objective function at each iteration requires only the off-line values \( \phi(u_j, w_j) \); the black-box model is never run along the search. This is a key feature in real-time applications where the stochastic optimization problem must be solved at different instances in time. If higher moments are needed to estimate, for instance, variance of the objective function then the design posterior distribution can be used. Note that reweighting can be used to compute confidence intervals for the gradient and thus have a more robust termination criterion. This comes, however, at a much higher computational cost because full problems must be solved with different realizations of uncertainty in order to compute the variance of the objective, at solutions satisfying the KKT conditions [20,22]. In the proposed approach, the search is terminated when the norm of the KKT residual is sufficiently small. This procedure will converge as long as the QP step obtained with the estimate \( \nabla_u \bar{V}(u^k) \) provides a direction of sufficient decrease on the actual objective \( V(u^k) \). This is normally observed in practice [23], and the chances are enhanced if sampling is performed along the search, as it is done here [24]. There is no general guarantee, however, that the solution is obtained unless an explicit inference task is performed. In the proposed approach, a cheap a posteriori validation is performed by computing the value of the stochastic solution to determine whether the procedure is improving the objective over the deterministic solution.

The framework for BONUS solver for this problem is shown in Fig. 2 where the WRF modeling framework is used to provide forecast of dry bulb temperature and humidity in terms of probability distributions. BONUS algorithm uses the sampling loop with the model to obtain initial probability distributions of objective function and constraints. This algorithm then uses reweighting scheme for optimization. The final output of this scheme is the optimal policy.

One of the limitations of this approach is that complexity increases with the number of degrees of freedom \( u \) and random variables \( w \) because the number of samples needed to cover the set \( \mathcal{U} \) and the uncertain space increases. Convergence is normally
exponential in the number of samples. Note, however, that this complexity is reflected in off-line computations only and can be reduced in practice by using variance reduction techniques such as Hammersley sequence sampling or Latin hypercube sampling [25]. The number of samples needed to estimate the distributions along the search by using KDE is problem dependent. Also note that the complexity of the QP increases cubically with the number of degrees of freedom since the QP it is nearly dense.

4. Weather forecasting and uncertainty quantification

This section presents the procedures used to compute real-time forecasts and uncertainty information for ambient temperature and humidity using the numerical weather prediction (NWP) model WRF. Details of the ensemble initialization and restarting procedures are provided. These procedures are capable of providing frequent forecast updates as needed in real-time power plant operations.

The WRF model [14] is a state-of-the-art numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs. WRF is the result of a multi-agency and university effort to build a highly parallelizable code that can run across scales ranging from large-eddy to global simulations. WRF has a comprehensive description of the atmospheric physics that includes cloud parameterization, land-surface models, atmosphere-ocean coupling, and broad radiation models. The terrain resolution can go up to 30 s of a degree (less than 1 km$^2$). The discretization of the underlying nonlinear partial-differential equation system over a 3-D field typically leads to a state dimensionality of O($10^3$–$10^6$). Consequently, the WRF model is computationally intensive, and special domain targeting techniques and computational resources are needed to enable computational feasibility [10]. An implementation of the WRF system at Argonne National Laboratory has been used. This implementation can be used to generate validated weather data and uncertainty information for energy optimization studies [15,26,10].

In order to initialize the NWP simulations, reanalyzed fields are used, that is, simulated atmospheric states reconciled with observations (the entire atmospheric state space is required by the model) as initial conditions, since only a small subset of the state space is available through measurement at any given time [27]. The reanalyzed states are generated from an internal data assimilation procedure. In particular, the North American Regional Reanalysis (NARR) data set is used to cover the North American continent (160W–20W; 10N–80N) with a resolution of 10 min of a degree, 29 pressure levels (1000–100 h Pa, excluding the surface), every 3 h from 1979 until present. An ensemble of realizations is used to represent uncertainty in the initial (random) wind field and propagate it through the WRF nonlinear model. The initial ensemble is obtained by sampling from an empirical distribution, a procedure similar to the National Centers for Environmental Prediction (NCEP) method [28]. More detailed explanations of the procedures needed are presented in the following subsections.

4.1. Ensemble initialization

In a normal operational mode, the NWP system evolves a given state from an initial time $t_0$ (current time) to a final time $t_f$. The initial state is produced from past simulations and reanalysis fields. Because of observation sparseness in the atmospheric field and the incomplete numerical representation of its dynamics, the initial states are not known exactly and can be represented only statistically. Therefore, a distribution of the initial conditions is used to describe the confidence in the knowledge of the initial state of the atmosphere. A normal distribution of the uncertainty field of the initial state is assumed, a typical assumption in weather forecasting. The distribution is centered on the NARR field at the initial time, the most accurate information available. In other words, the expectation is exactly the NARR solution. The second statistical moment of the distribution described by the covariance matrix $\Sigma$ is approximated by the sample variance or pointwise uncertainty. The initial $N_s$-member ensemble field $\chi_i^{0} := \chi_i(t_0), \; i \in \{1 \ldots N_s\}$, is sampled from $\mathcal{N}(\chi_{\text{NARR}}; \Sigma)$:

$$\chi_i^{0} = \chi_{\text{NARR}} + \Sigma^{1/2} \xi_i, \; \xi_i \sim \mathcal{N}(0, I), \; i \in \{1 \ldots N_s\}. \tag{4.18}$$

This is equivalent to perturbing the NARR field with $\mathcal{N}(0, \Sigma)$. That is, $\chi_j = \chi_{\text{NARR}} + \mathcal{N}(0, \Sigma)$. Note that the correlation matrix is huge (square of the discretized number of states in a 3-D field) and thus cannot be formed and stored. In what follows, the procedure used to estimate the state correlation matrix is described.

4.2. Estimation of the correlation matrix

In weather models the correlation structure typically is localized in space. Therefore, in creating the initial ensemble one needs to estimate the spatial scales associated with each variable. To obtain these spatial scales, correlation matrices of the forecast errors are built using the WRF model. These forecast errors are estimated by using the NCEP method [28], which is based on starting several simulations staggered in time in such a way that, at any time, two forecasts are available. In particular, day-long simulations are run over a month. These runs are started every 12 h so that every 12 h two forecasts are obtained, one started 1 day before and one started half a day before. The differences between two staggered simulations is denoted as $d_i \in \mathbb{R}^{30 \times (2 \times 30 \text{ days})}$, that is, the difference at the $i$th point in space between the $j$th pair of forecasts, where $N$ is the number of points in space multiplied by the number of variables of interest. Symbol $e_i$ defines the $i$th row, each of which corresponds to the deviations for a single point in space. Therefore, the covariance matrix can be approximated by $\Sigma \approx \text{dd}^T$. Calculating and storing the entire covariance matrix are computationally intractable. Consequently, the correlation distance at each vertical level and for each variable by two parameters representing the East–West and North–South directions is considered. This approach captures the Coriolis effect and the Earth rotation, as well as faster and larger-scale winds in the upper atmosphere. Correlations and winds are assumed to be roughly similar in nature across the continental U.S. This process is repeated in the vertical direction. To create the perturbations from these length scales, a normally distributed noisy field is used, and Gaussian filters are applied in each direction with appropriate length scales to obtain the same effect as in (4.18).

4.3. Ensemble propagation through the WRF model

The initial state distribution is evolved through the WRF model dynamics. The resulting trajectories can then be assembled to obtain an approximation of the forecast covariance matrix:

$$\chi_i^{0} = M_{t_0 \rightarrow t} \cdot (\chi_i^{10} + \eta_i(t), \; s \in \{1 \ldots N_s\}, \tag{4.19}$$

where $\chi_i^{10} \sim \mathcal{N}(\chi_{\text{NARR}}, \Sigma), \eta_i \sim \mathcal{N}(0, I)$, and $M_{t_0 \rightarrow t}$ represents the evolution of the initial condition through the WRF model from time $t_0$ to time $t_f$. The initial condition is perturbed by the additive noise $\eta$ that accounts for the various error sources during the model evolution. An analysis of the covariance propagation through the model is given in [10].

In this study, it is assumed that the numerical model (WRF) is perfect, that is, $\eta = 0$. Thus, given the exact real initial conditions, the model produces error-free forecasts. For long prediction windows; this is a strong assumption. In this study, however, the
forecast windows are restricted to be no longer than 1 day ahead, thus making this assumption reasonable.

Relative humidity (RH) is typically a derived quantity or a diagnostic variable in the NWP model, which means it is not modeled directly. RH is computed based on air density, temperature, and vapor pressure [14]. In particular, the water mixing ratio is available as a standard WRF output; then the RH is obtained as the ratio between the water mixing ratio and the saturation mixing ratios, which are based on local temperature and pressure.

5. Integration study

A numerical study is presented in this section. The study integrates stochastic optimization and weather uncertainty quantification using detailed physics models for weather and power plant simulation. The framework is sketched in Fig. 3. The results show that stochastic optimization can yield significant improvements over deterministic operating policies. In addition, the computational feasibility of the proposed framework is demonstrated.

5.1. Plant description

A general scheme of a pulverized coal power plant is shown in Fig. 4. A more detailed description can be found in [12,13,29] and the references therein. Powdered coal is blown into the boiler combustion chamber, and the hot gases and heat produced by the combustion of coal are used to produce steam within the tubes inside the boiler. High-pressure steam is fed into a train of high pressure (HP), intermediate-pressure (IP) and low-pressure (LP) turbines where it does mechanical work causing the turbine shaft to rotate at high speed. The steam returns to the boiler after the first pass through HP turbines, where it is reheated before being fed to the IP turbines. After passing through the turbine train, the steam is cooled to condensation and returned to the boiler while being preheated with steam extracted from the turbines in eight stages.

A flue gas desulfurization (FGD) unit is attached to the boiler effluent to remove the SO₂ before the effluent is released to the atmosphere. This FGD unit employs limestone and oxygen in a reactive absorber to react with SO₂ and produce calcium sulfate (gypsum). Limestone is fed to the unit as slurry, and the resulting sulfate is removed from the absorber as suspension water. In order to compensate for losses by evaporative cooling and water employed in the flue gas treatment units (ash removal and flue gas desulfurization) large makeup streams are required in the cooling cycle. As previously discussed, the amount of make-up water can become constrained under different scenarios, thus limiting capacity.

The model is a supercritical steady-state flow sheet without carbon capture designed to generate 700 MW of electricity. Carbon capture technologies are water intensive. The amine based technology currently under consideration for commercialization imposes water penalty as much as 50% in a coal fired power plant [3]. The process performance parameters (generation, efficiency, emission and water consumption) are calculated with the process simulator model Aspen Plus, which allows the use of available unit operation models, property databases, and simulation algorithms. The specifications of the Aspen model are summarized in Table 1. A sequential modular approach was used to simulate the model.

Accurate estimation of the water intake strongly depends on the cooling tower model used. In this work, an equilibrium-based model for the cooling tower was used to estimate the evaporation rates [12]. The model comprises two flash units and one heat exchanger. One of the flashes is set to be adiabatic and is employed to calculate the wet bulb temperature from dry bulb temperature and relative humidity. The second flash is coupled with the heat exchanger and has no liquid outlet. A design specification ensures that heat provided to the flash equals the condenser cooling requirement. Additionally, a set of design specification and calculators determine the cold water temperature, circulating water flow rate, and air flow rate for a constant volume, forced-drift cooling tower. A constraint on the maximum water intake of $9.1 \times 10^5$ kg/h is enforced. This

---

**Fig. 3.** Integration framework of weather uncertainty quantification and stochastic optimization.
represents the nominal consumption under average atmospheric conditions: of 288 K and 60% relative humidity.

In order to determine the plant operational and design parameters that influence water consumption the most and therefore the maximum generation, a partial rank correlation coefficients analysis is performed. Reheater temperature, air excess to the boiler, water content of the FGD slurry, and pressure drops at HP turbines 1 and 2 were chosen as decision variables (degrees of freedom).

5.2. Validation of weather forecast

Figs. 5 and 6 present average and 30 ensemble profiles are presented for the dry bulb temperature and relative humidity for a day, respectively, in a random location in the U.S. Midwest region for June 1, 2006. Dots are real measurements from meteorological stations. The day-long profile is forecast at the location of a meteorological tower located near Chicago, IL (41°42′04 ″, 87°59′42 ″). The crossed dots show the actual temperatures and humidity observed at the closest meteorological station as measured by the instruments mounted on the weather tower. The envelope surrounding the ensembles is the 95% confidence region. Time zero represents 6 a.m. central time.

One can see that the projection for a day ahead is reasonably accurate from the WRF model, capturing well the trend and sensor readings. Temperature varies significantly throughout the day in the range of 288–299 K, while the relative humidity varies in the range of 20–90%. Humidity is much higher at night, whereas temperature is much higher during the day, the variables are anticorrelated. The temperature and RH uncertainties are the result of different NWP model forecasts for the same prediction window. While the temperature is relatively well modeled, the RH is the result of the moisture transport in the atmosphere and complex physical interactions that lead to phase changes. Therefore, as expected, a relatively wider uncertainty estimate for RH and a fanning effect toward the end of the forecast window are observed. It should also be noted that, while uncertainty for temperature projections has been discussed in other studies, projected uncer-

---

Table 1

<table>
<thead>
<tr>
<th>Units</th>
<th>Number of units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sections</td>
<td>3</td>
</tr>
<tr>
<td>Boiler</td>
<td>1</td>
</tr>
<tr>
<td>Steam cycle</td>
<td>1</td>
</tr>
<tr>
<td>Cooling water</td>
<td>1</td>
</tr>
<tr>
<td>Total streams</td>
<td>178</td>
</tr>
<tr>
<td>Mass streams</td>
<td>120</td>
</tr>
<tr>
<td>Energy streams</td>
<td>37</td>
</tr>
<tr>
<td>Work Streams</td>
<td>21</td>
</tr>
<tr>
<td>Total ASPEN blocks</td>
<td>128</td>
</tr>
<tr>
<td>Unit operations</td>
<td>115</td>
</tr>
<tr>
<td>Heat mixers/splitters</td>
<td>10</td>
</tr>
<tr>
<td>Work mixers/splitters</td>
<td>3</td>
</tr>
<tr>
<td>Calculator blocks</td>
<td>25</td>
</tr>
<tr>
<td>Design specifications</td>
<td>31</td>
</tr>
<tr>
<td>Convergence blocks</td>
<td>31</td>
</tr>
</tbody>
</table>

---

Fig. 4. Schematic representation of coal-fired plant.

Fig. 5. Forecast (thick line) and ensemble profiles (thin lines) for dry bulb temperature.
The uncertainty envelopes for the ambient temperature are narrow. The uncertainty envelope for the relative humidity, on the other hand, is wide and can be as large as 30–40% toward the end of the day. Both envelopes follow complex shapes because of the nonlinearity of the weather prediction model and current meteorological conditions. Toward the end of the day, the mean forecast does not predict well the shape of the humidity profile, but the uncertainty envelope covers this region reasonably well. These results indicate the additional complexity in forecasting humidity. They also reflect the fact that the uncertainty envelope is unable to fully cover the sensor readings at all times.

The ambient temperature and relative humidity WRF ensembles are fitted to a log-normal and uniform distributions, respectively. These are illustrated in Figs. 7 and 8. These were used to generate the design distribution \( P_d(u, w) \) and corresponding objective function posterior \( P_d(\varphi(u, w)) \).

### 5.3. Maximum power performance

In order to demonstrate the economic benefits of stochastic optimization over current practice, extensive simulations using three settings were performed. In the first setting, the here-and-now stochastic optimization problem was solved to maximize the expected value of power during a timeframe of 24 h with time steps of 1 h. The WRF ensemble information was used to solve the problem. This profile is labeled as “Stochastic”. In the second setting, the deterministic problem was solved using the mean forecast obtained from WRF and the expected value of the maximum power was computed using the ensembles from WRF. The maximum power profile is labeled “Deterministic (b)”. In the third setting, day-average conditions for temperature and humidity are computed. In addition, the operating conditions throughout the day are fixed. The expected value of the maximum power was again computed from the WRF ensembles. The maximum power profile is labeled “Deterministic (a)”. Fig. 9 shows the result of optimization. The following findings were obtained:

- From Deterministic (a) it is observed that the power profile does vary with the weather conditions throughout the day. Specifically, the power levels vary from 715 to 665 MW, which represent 7% of the base load of 700 MW. Consequently, the setpoints must be adjusted throughout the day to mitigate the effects of weather and water constraints.
- High power is obtained at night when ambient temperature is lower, even if humidity is high. Low power is obtained around noon when temperature is high, even if humidity is low. These results indicate that temperature dominates the cooling capacity. In addition, it indicates that the plant becomes constrained by cooling capacity at high temperatures since a constraint on
water intake is imposed.
- From Deterministic (a) and (b) it can be seen that adjusting operating conditions according to the mean forecast can increase the maximum power output. The total output for (a) is 16,675 MW h, while for (b) it is 16,823 MW h, a performance gain of 1.46 MW h.
- Stochastic optimization gives a total output of 16,922 MW h, a performance gain of 99 MW h over Deterministic (b) and 245 MW h over (a).
- Stochastic optimization displaces the power output curve but does not fully mitigate the variability of power throughout the day. In particular, the variability for the stochastic profile is 45 MW. This indicates the strong impact of weather and corroborates the fact that the plant is constrained by cooling capacity.
- At hour 15 and 23 the stochastic profile is lower than that of Deterministic b), because of the estimation error introduced from the small number of samples. The profiles, however, indicate that the stochastic optimization solution is consistently better than the deterministic one, indicating that a small number of samples seem sufficient to obtain accurate solutions. This, in turn, indicates that the weather uncertainty information is accurate and the reweighting scheme is efficient in identifying good-quality solutions.

These results indicate that, for a given water intake constraint, power can be maximized by using stochastic optimization. They also imply that, as water constraints are tightened, it is possible to maintain high output levels with stochastic optimization, thus reducing water consumption. Note that tightening the water intake constraint will increase the sensitivity of the power profile to weather conditions.

If the gain of 245 MW h is projected over a total of 100 days in a year (summer conditions), this translates into 24.5 GW h that, if paid at a price of $100/MWh, translates into $2,450,000. Of course, a more rigorous analysis capturing seasonal effects and different constraint scenarios is needed. The presented analysis, however, provides an order of magnitude of the expected savings through stochastic optimization.

The potential of stochastic optimization in aiding bidding procedures should be highlighted. In particular, the proposed approach enables a more robust prediction of the expected performance of the plant. This is important in day-ahead bidding procedures where maximum capacity is an important parameter needed in market participation. The computed maximum capacity profiles obtained with stochastic optimization can implicitly minimize economic penalties arising in real-time markets. In particular, if capacity is overestimated, then the plant owner will have to pay for the unmet generation at a real-time price, which tends to be higher because more expensive peaking plants are used in real-time grid operations. On the other hand, if capacity is underestimated, there will be an opportunity cost if the available surplus cannot be sold in the real-time market.

5.4. Numerical performance

A single simulation of the Aspen power plant model takes, on average, 10 min on a Intel Core 2 Duo processor running at 2 GHz. This time is required for all the loops to converge and satisfy the design specifications within the simulator. In order to set-up the initial distributions of the objective function, 600 scenarios were run. Each scenario required a simulation of the Aspen model (10 min). This required a total off-line time of about 100 h. To generate these scenarios, the uncertain variable domain was sampled (from uniform distributions) considering weather variations of 1 week. Note that the design distributions can be reused to solve stochastic optimization problems over several days. This can be done while updating the design distributions for the next week as new weather data is obtained.

The average number of SQP iterations was 7, with a minimum of 2 and a maximum of 16. Each SQP iteration takes less than a minute, and 50 different samples were used at each SQP step. To verify that global optimality was achieved, the optimizer was run from 7 different starting points at each time step. A total of 1184 SQP iterations was required for the 24-h timeframe. Starting from multiple starting points also enhances the chances of identifying higher-quality solutions [24].

A traditional stochastic optimization approach (with no reweighting) would have used 355,200 runs of the Aspen model to solve the problems over the 24-h timeframe. That in turn would have required 59,200 h of computation (2400 days). This total number of runs comprises 6 Aspen model simulations per SQP iteration (one for function evaluation and one for each degree of freedom to evaluate derivatives). The proposed solution approach with reweighting does not require on-line model runs. The total on-line time for 1184 SQP iterations is around 2 h. The proposed solution offers 99.8% savings in computational load and enables feasibility.

The strategy for estimating the uncertainty in weather forecasts is highly parallelizable because each scenario evolves independently through the numerical weather model WRF, once the initial ensemble has been generated. The most expensive computational element is the evolution of each member through the WRF system. A two-level parallel implementation scheme is considered. The first level is a coarse-grained task division represented by each individual member. A secondary fine-grained level consists in the parallelization of each individual member. This approach is readily available because WRF is already parallelized, and this strategy yields a highly scalable solution.

The WRF simulations were carried out on Argonne National Laboratory’s Fusion cluster. This cluster has 320 nodes each with two Nehalem 2.6 GHz Pentium Xeon and 36 GB of RAM. Three nested domains were considered, the coarsest having a horizontal resolution of 32 by 32 km (#1), and the finest 2 by 2 km (#3), as illustrated in Fig. 10. This targeted domain partition provides an accurate representation of the target region, which was selected to be the state of Illinois, and dramatically decreases computational complexity. In particular, it avoids the use of fine uniform resolutions over the entire country, which is the approach used by most providers of forecast information.

For this study, NWP simulations were carried out asynchronously on a small number of cores. It is estimated, however, that members for a 24-h simulation can be computed in less than an hour on about 400 cores. Consequently, it is feasible to perform

![Fig. 10. Simulation domain and resolution for WRF.](image-url)
rigorous uncertainty quantification tasks with WRF to aid power plant operations.

6. Conclusions and future work

A stochastic optimization framework is proposed for performing integrated water and energy management in cooling-constrained power plants. The approach integrates detailed physical weather and power plant simulation models to assess effects of ambient conditions on plant performance. The results show that the maximum plant capacity can vary in the range of 5%–10% of the nominal capacity in intraday operations because of weather variations. In addition, it has been found that stochastic optimization can lead to daily capacity gains of as much as 245 MW h over current practice and enables more robust bidding procedures. In addition, a reweighting scheme makes the approach computationally feasible.

The results open interesting directions for future work. A more extensive study is needed of the effects of seasonal weather conditions on plant performance and on the accuracy of the weather forecasts throughout the year. This will require, however, extensive computational resources to perform integrated weather forecasting and stochastic optimization. It is also necessary to revisit the internal reweighting scheme to assess, on the fly, the quality of the solutions using inference techniques. Also of interest would be a study of the effects of weather on other plant configurations such as combined cycle power plants where ambient conditions affect combustion processes. Formulations using dynamic effects such as ramp constraints must also be developed; these will increase dramatically the complexity of the problem, since time decoupling will no longer be possible. Furthermore, it is of interest to study the effect of risk functions and long-term water reservoir dynamics (e.g., seasonal) on plant performance.

Acknowledgments

This work was supported by the U.S. Department of Energy, under Contract No. DE-AC02-06CH11357. The authors thank Argonne National Laboratory’s Laboratory Computing Resource Center for use of the Fusion cluster.

References