Xiaosheng Zhang, *Student Member, IEEE*, Tao Ding, *Senior Member, IEEE*,Chenggang Mu*, Student Member, IEEE,* Biyuan Zhang*, Student Member, IEEE,* Yuankang He, Mohammad Shahidehpour, *Fellow*, *IEEE*

Data, Parameters, and Training Procedure

The simplified topology of a regional power grid in China is shown in Fig. 1. This system is a high-penetrated renewable power system. The capacity of conventional thermal plants, hydro plants, and RES plants is shown in Table I. The scenario tree is constructed by using the Latin Hypercube Sampling method. Fig. 2 plots the forecasted RES output and gross load profiles of the regional power grid. The maximum and minimum load demands are 95,000 MW and 83,500 MW.

The parameters of the adding VESs are listed in Table II.



Fig. 1 Simplified topology of a regional power grid in China.

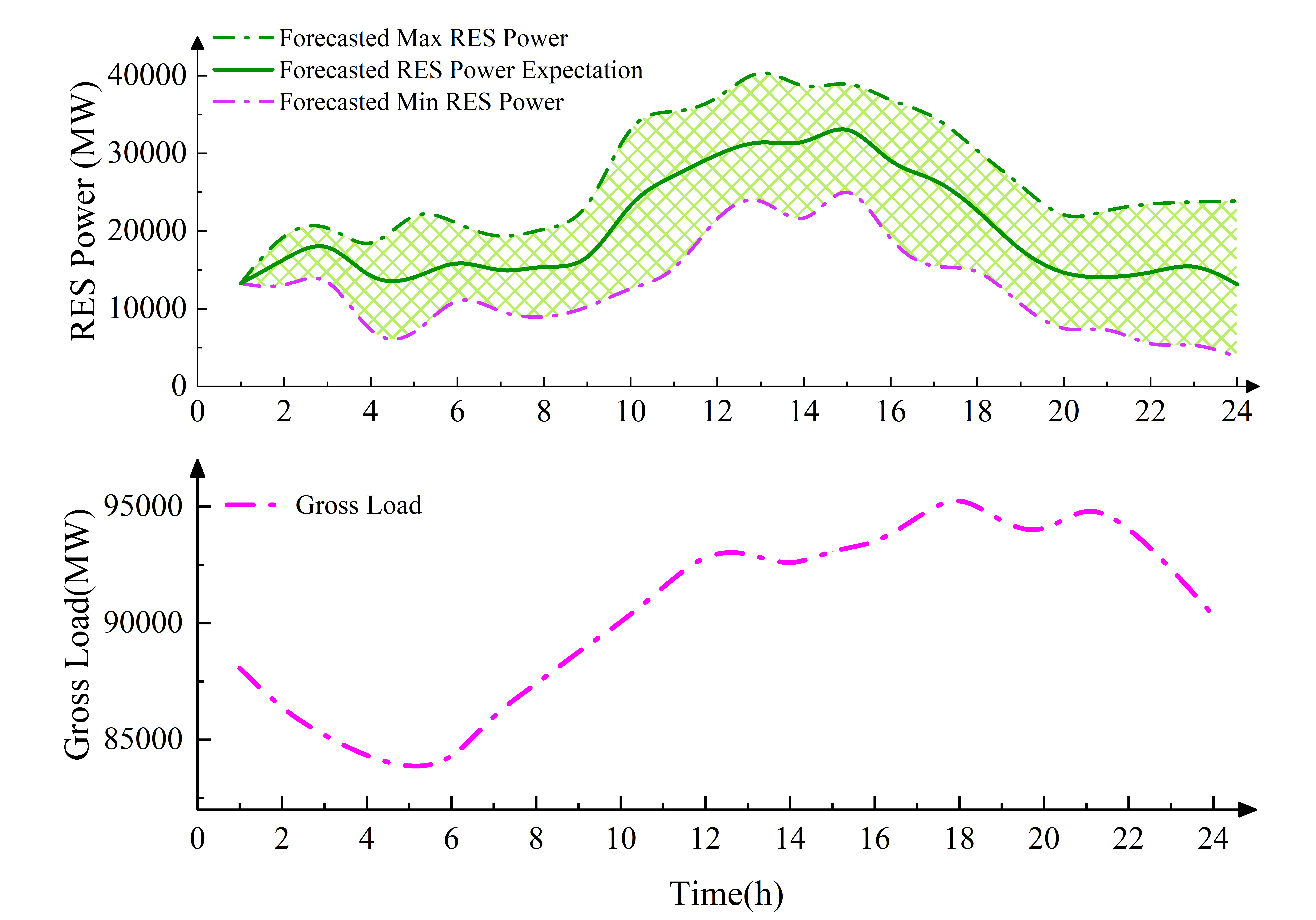


Fig. 2 (a) The forecasted RES output and (b) Gross load of the regional power grid

Table I. Capacity of Different Types of Units (MW)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Units | Thermal | Hydro | RES | Total |
| Capacity(MW) | 102643 | 31323 | 60738 | 194704 |

Table II Parameters of the VES

|  |  |  |  |
| --- | --- | --- | --- |
| *g* | Bus | PG j,max/PG j,min (MW) | up *j*/down *j* |
| 1 | 5 | 1200/600 | 300 |
| 2 | 16 | 2400/1200 | 600 |
| 3 | 24 | 2400/1200 | 600 |
| 4 | 39 | 1200/600 | 300 |
| 5 | 48 | 1920/960 | 480 |
| 6 | 140 | 2400/1200 | 600 |
| 7 | 168 | 900/450 | 225 |
| 8 | 169 | 2400/1200 | 600 |

## Training Procedure

In the related literature, the above problem is usually solved by gradient-descend type algorithms, while the parameter ** is regarded as the variable. The advantage of gradient-descend type algorithms is that they do not require any extra information in addition to the input and output values of the function. In our problem, ** t(at-1) is convex, and the sub-gradient of ** t(at-1) is derived by solving the linear programming. Thus, the problem can be solved more efficiently with the above information. By regarding ******1,..., ** as constant, the optimality condition of (37) in the paper is stated as,



Due to the convexity of ** t(at-1), we have



Combining and , V** t(at-1; **) is stated as follows,





where ** is the intercept of the cut and  is the gradient of the cut. SDDPIL is a model-based reinforcement learning algorithm since the environment model p(st+1|st, at) is known as p(|). It is a double-pass reinforcement learning algorithm consisting of a forward pass and a backward pass. Accordingly, the detailed training procedure is stated as follows.

### Forward Pass

The forward pass will solve the Bellman optimality equation by outer approximating the expected cost-to-go function ** t+1(at) with piecewise linear functions at each stage. The dynamic programming equation for t = 1, 2,…, T is stated as follows



where V** t+1 is the piecewise linear approximations of ** t+1(at). The detailed formulation is stated as follows,



where  is the intercept of cut e;  is the gradient of cut e; t+1 is the set of the cuts in stage t+1.

The task of the forward pass is to generate trajectories for updating V** t+1 in the backward pass. The forward pass uses a Monte Carlo sampling technique to generate  scenarios  from the scenario tree. Then will be solved on the  scenarios from t = 1, 2,…, T to generate  trajectories. For stage T, we have V** T+1(aT) ≡ 0.

### Backward Pass

The task of the backward pass is to generate supporting hyperplanes to update V** t. The backward pass will solve from t = T to t = 2 with the trajectories extracted randomly from the experience replay buffer. After solving for all scenarios at stage t, the dual variables **t of the constraint (st) = () and the optimal values  (st ()) are obtained to generate the cuts, which are added to all sub-problems at stage t-1 to update V** t. The gradient  and intercept  of cut e are updated as follows:





The cuts will update the V** t to improve the approximation of the expected cost-to-go function iteratively. At the early iterations of training, the expert performance data are better than the trajectories generated in the forward pass. The expert performance data are more likely to be chosen by giving a larger probability. After several iterations, the expert performance data will be given the same probability as the trajectories generated in the forward pass. This means that in later iterations, the forward passing trajectory will be treated as important as expert performances.

## Convergence Criterion of SDDPIL

Different from model-free methods, SDDPIL has an explicit convergence criterion which is guaranteed by the probability theory. The SDDPIL provides the lower bound (LB) by solving the first stage problem.



The upper bound (UB) is provided by calculating the expectation of the return of several episodes. The UB is a confidence interval for the random sampling.







where z*α*/2 is the (1 – *α*) quantile of the standard normal distribution. Finally, the flowchart of the SDDPIL algorithm is shown in Table I.