Available online at www.sciencedirect.com
SCIENCE DIRECT
Energy and Buildings 38 (2006) 142–147
ENERGY and BUILDINGS
www.elsevier.com/locate/enbuild

Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory Part 1. Theoretical foundation

Simeng Liu [a,*], Gregor P. Henze [b]

[a] Architectural Engineering, University of Nebraska-Lincoln, 1110 South 67th Street, PKI 243, Omaha, NE 68182-0681, USA
[b] Architectural Engineering, University of Nebraska-Lincoln, 1110 South 67th Street, PKI 203D, Omaha, NE 68182-0681, USA

Received 31 March 2005; received in revised form 28 May 2005; accepted 3 June 2005

Abstract

This paper is the first part of a two-part investigation of a novel approach to optimally control commercial building passive and active thermal storage inventory. The proposed building control approach is based on simulated reinforcement learning, which is a hybrid control scheme that combines features of model-based optimal control and model-free learning control. An experimental study was carried out to analyze the performance of a hybrid controller installed in a full-scale laboratory facility. The first part presents an overview of the project with an emphasis on the theoretical foundation. The motivation of the research will be introduced first, followed by a review of past work. A brief introduction of the theory is provided including classic reinforcement learning and its variation, so-called simulated reinforcement learning, which constitutes the basic architecture of the hybrid learning controller. A detailed discussion of the experimental results will be presented in the companion paper.

Keywords: Load shifting; Thermal Energy Storage (TES); Optimal control; Learning control; Reinforcement learning

1. Introduction

1.1. Motivation

The advantages of shifting building cooling load by using active and passive thermal storage capacity have been realized for a long time. By definition, *active* building thermal capacity refers to a thermal energy storage (TES) system, which is either a chilled water or ice based system. *Passive* building thermal storage capacity refers to the building envelope, internal construction and furniture, which affect the building cooling load. The motivation for this study stems from a research project that investigates predictive optimal control of active and passive building thermal storage inventory. In this project, a model-based predictive optimal controller was developed in order to evaluate the merits of controlling active and passive thermal storage optimally in a continuous closed-loop fashion. The first two phases of the study, numerical analysis and experimentation, demonstrated that an optimal controller can operate the building and cooling plant efficiently and achieve significant cost savings. However, the model-based predictive optimal control approach is strongly affected by the quality of the model. Efforts to develop and maintain the model can be very demanding, time-consuming, and costly.

To overcome these shortcomings, a new methodology, reinforcement learning control, was introduced as the third phase of the study to tackle this control problem in an adaptive manner. Results of a numerical analysis and simulation study show that although pure reinforcement learning can direct the controller to approach a near-optimal control strategy, it usually takes an unacceptably long time to make the controller ''learn''. The slow nature of reinforcement learning makes it almost impossible to implement such control algorithm directly into any practical application.

DOI of original article: 10.1016/j.enbuild.2005.06.001.
* Corresponding author.
E-mail addresses: sliu@mail.unomaha.edu (S. Liu), ghenze@unl.edu (G.P. Henze).
URL: http://www.ae.unomaha.edu/ghenze

0378-7788/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.enbuild.2005.06.002

As a result, a hybrid control scheme has been proposed that attempts to combine the positive features of both the model-based and reinforcement learning approach. The hybrid control approach is based on a variation of classic reinforcement learning called *simulated reinforcement learning*, which inherits the basic structure of reinforcement learning. However, the controller is trained by a model in a simulation environment before it is implemented in a real control application. In order to validate the hybrid control approach, experimentation was carried out in the same laboratory facility as the model-based approach experiment. This paper summarizes the findings of the study leading to the development of the novel hybrid control approach and presents selected experimental results. The investigation shows that although the proposed hybrid control approach functions as expected, it is affected by the training model and other learning parameters.

1.2. Review of past work

Previous studies on building thermal mass utilization demonstrate the potential of reducing peak cooling loads and associated electrical demand. The results show that cost savings vary widely among the published case studies [1–4]. In a simulation study presented by Braun [5], cost savings for a design day varied from 0 to 35% depending on system type and utility rate. Andresen and Brandemuehl [6] showed energy and cost savings potential by precooling the building structure, calling attention to the importance of the mass of furnishings that significantly affects the precooling strategy. In a review article on load control using building thermal mass, Braun [7] concluded that the savings potential is very sensitive to the utility rates, building and plant characteristics, and weather conditions and occupancy schedule. The greatest cost savings were realized for the case of heavy construction, good part-load characteristics and low ambient temperature that enabled free cooling during nighttime ventilation.

Optimal control of TES has been investigated by several researchers. To evaluate the theoretical potential of ice storage systems in reducing operating costs, a detailed analysis was performed by Henze et al. [8,9] using a dynamic programming-based simulation environment. Within this environment, a set of three conventional control strategies was compared to optimal control, which in turn served as a benchmark to determine how well conventional controls harnessed the system's cost saving potential. A subsequent investigation phase presented in companion papers by Henze et al. [10] and Henze and Krarti [11] determined to what extent the performance merits of optimal control are retained when the optimal controller is subjected to uncertainty in the external variables influencing the physical process, such as future weather variables and cooling loads.

The project *Predictive optimal control of active and passive building thermal storage inventory*, sponsored by the U.S. Department of Energy, attempts to combine these merits and mathematically analyze the optimization at the same time. A simulation study was carried out to investigate the combined usage of active and passive building thermal storage inventory by Henze et al. [12]. The analysis showed that when an optimal controller for combined utilization is given perfect weather forecasts and when the building model used in the model-based predictive control perfectly matches the actual building, the utility cost savings are significantly greater than either storage, but less than the sum of the individual savings. In addition, the cooling-on-peak electrical demand can be drastically reduced. Further research by Henze et al. [13] also demonstrated that prediction uncertainty in the required short-term weather forecasts can affect the controller's cost saving performance. Liu and Henze [14] investigated the impact of five categories of building modeling mismatch on the performance of model-based predictive optimal control of combined thermal storage using perfect prediction. The results showed that a simplification or mismatch of the building geometry and zoning only marginally affected the optimization strategy. However, the mismatch of internal heat gain, building construction and energy system efficiency can lead to significant deviations in the optimization. Henze et al. [15] demonstrated model-based predictive optimal control of active and passive building thermal storage inventory in a test facility in real time using time-of-use differentiated electricity prices without demand charges. The experiment essentially confirms the previous findings in the numerical analysis of optimal control of building thermal storage. However, the savings associated with passive building thermal storage inventory proved to be small because the test facility is not an ideal candidate for the investigated control technology.

The learning control approach was considered when dependence on model quality became evident in the model-based approach. Earlier research by Henze and Dodier [16] investigated learning control of a grid-independent photovoltaic system consisting of a collector, storage, and a load. Better performance was found by applying reinforcement learning to optimize the operation of the system. Henze and Schoenmann [17] investigated the application of reinforcement learning control to the optimization of a thermal energy storage system. Although reinforcement learning control proved sensitive to the selection of state variables, level of discretization, and learning rate, the controller effectively learned how to control thermal energy storage and displayed good performance. The cost savings compared favorably with conventional cool storage control strategies, but did not reach the level of predictive optimal control.

These studies encouraged the authors to investigate reinforcement learning for optimal control of active and passive building thermal storage inventory. In a follow-up research study, a simulation environment was developed, in which a supervisory controller was designed to control the

thermal storage inventories of a commercial building model based on a reinforcement learning algorithm. Liu and Henze [18,19] demonstrated that the reinforcement learning approach can find the optimal or a near-optimal control policy without prior knowledge of the environment, but it takes an unacceptably long time. Furthermore, the performance of the controller was sensitive to many factors including selection of the state-action space and learning parameters. Implementation of such a controller with no prior domain knowledge would not be practical in any real building control application.

As a result, a hybrid control scheme is proposed that combines the merits of the model-based approach and the model-free learning approach. The hybrid approach is based on a variation of the classic reinforcement learning approach and is called *simulated reinforcement learning*. The following sections describe the efforts made to develop the hybrid learning controller. The methodology behind the approach and an analysis of the hybrid control experimental study are presented, and the performance of the controller is compared with model-based optimal control and other conventional control strategies.

2. Development of a hybrid control approach

2.1. Problem statement

Our problem can be formulated as a sequential decision-making problem, in which an intelligent controller is continuously facing a situation that requires it to select an action a ($a \in A$) when the condition of the environment is at a specific state s ($s \in S$), in order to maximize cumulative rewards $\sum r_t$ given a time horizon T ($t = 1, 2, \ldots, T$). The strategy to select the certain action at a given state is called policy $\pi(s, a)$, and our goal is to find the optimal policy $\pi^*(s, a)$ that leads the controller to select an appropriate action a in any given state s to maximize the cumulative rewards. Specifically, the desired controller tries to minimize a cost function:

$$J = J(u_1^*, \ldots, u_l^*) = \min\left[\sum_{k=0}^{l} r_k P_k \Delta t\right] \tag{1}$$

Eq. (1) describes a building HVAC controller that performs a sequence of actions over a selected time horizon of l to minimize the total cost at the end of period. In this equation, r_k is the price of electricity at time k. P_k is the total building electricity consumption, which is either the sum of cooling and non-cooling electrical loads, or the cooling-related load only. Δt is the time interval. In our case, the actions are specifically defined as the building global zone air temperature setpoint T_{sp}, the control variable that exploits the passive thermal storage inventory, and a control command for the active thermal storage system that is either charge (+) or discharge (−) rate u. The dynamics of a simple ideal TES model (without considering transmission heat losses) can be expressed in the following equation:

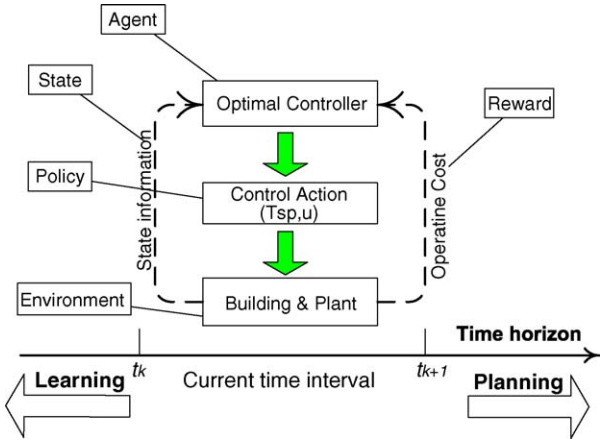

Fig. 1. Controller structure.

$$x_{k+1} = x_k + u_k \frac{\Delta t}{\mathrm{SCAP}} \tag{2}$$

where x_k is the state-of-charge of the TES, and SCAP is the TES capacity. Both the state-of-charge x and the control variable u are subjected to time-dependent constraints:

$$x_{\min} \le x \le x_{\max} \tag{3}$$

$$u_{k,\min} \le u_k \le u_{k,\max} \tag{4}$$

Fig. 1 depicts the overall structure of the controller. At the beginning of each time interval, the controller will first sense the current information of the building and plant, and then carry out the optimization based on prediction and planning, knowledge learned from the past, or a hybrid control algorithm that combines both features. Once the optimal control actions are found, the commands (including zone air temperature setpoint T_{sp} and the TES charge or discharge rate u) will be sent to the building automation system (BAS). As the actions selected by the controller are executed by the building and the plant, an operating cost will be generated at the end of the time interval. This information will also be fed back to the controller as the reward information of the selected action. The goal of the controller is to develop a strategy to find the optimal actions that minimize the cumulative operating costs over the total time horizon.

2.2. Methodology

A sequential decision-making problem can be solved by either *planning* or *learning* depending on the availability of knowledge of the environment. In planning problems, we assume that a complete model of the environment is available in advance. The task of finding the optimal policy is accomplished by a planner, and the optimal policy is then executed by the agent. On the other hand, when the environment is unknown or it is difficult to develop a

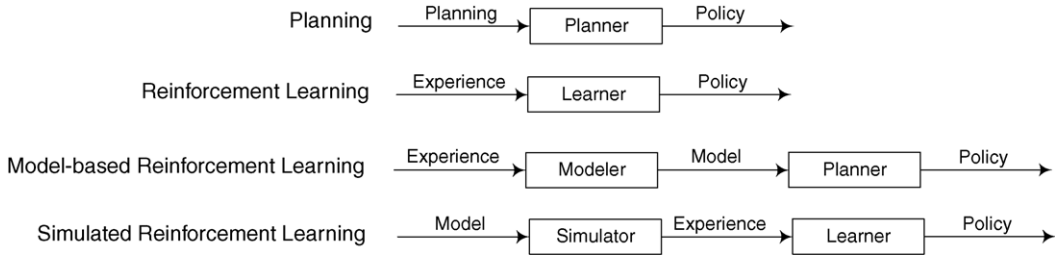

Fig. 2. Classification of sequential decision-making problems.

detailed model of the environment, the only way for the agent to find the optimal policy is through interaction with the environment including perception and action. This is the so-called reinforcement learning problem.

Fig. 2 presents four categories of sequential decision-making problem scenarios classified by Littman [20]. Besides the planning and reinforcement learning approaches introduced above, there are two additional cases that combine some characteristics of planning and learning. Model-based reinforcement learning uses the experience to generate a model, which serves as a planner to update the policy for the agent. On the other hand, in simulated reinforcement learning, a simulator is constructed first without taking the actual response of the system into consideration. The model can be developed, based on previous knowledge, which emulates the environment and generates the quasi-experience to train the learning controller. Next, the controller is implemented into an actual building application to direct the system by using the trained knowledge. This is the basic framework for the hybrid controller discussed in this paper.

2.2.1. Classic reinforcement learning

As mentioned earlier, the proposed hybrid learning control scheme is based on a variation of the classic reinforcement learning algorithm. It is necessary to introduce briefly the background of reinforcement learning before jumping into the discussion of the hybrid control scenario.

Fig. 3 sketches the general layout of a typical reinforcement learning problem. As depicted in Fig. 3, at any moment t, the agent first senses the current condition of the environment, which is represented as state S_t, and then selects an action a_t. The action causes the environment to change to a new state S_{t+1}. After the state transition, the agent receives a reward r_t. In a reinforcement learning problem, an agent interacts with the external environment to achieve a long-term goal, which is a measure of cumulative rewards over a finite or infinite sequence of decisions.

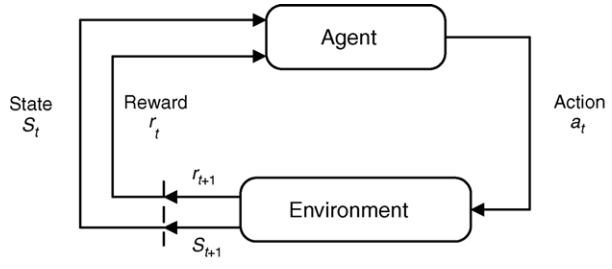

Fig. 3. Schematic of the reinforcement learning problem.

Most reinforcement learning control problems adopt the framework of a Markov decision process (MDP) that exhibits the Markov property [21]. A process is Markovian if the next state of the environment depends and only depends on the current state and current action to take. This property does not mean that the historical states are not important, only that all the historical information can be retained by the current state. In this case, a transition probability function is then introduced, defined as

$$p^a_{ss'} = \Pr(s_{t+1} = s' | s_t = s, a_t = a) \tag{5}$$

Similarly, the next reward is also defined as a function of current state, current action, and next state:

$$R^a_{ss'} = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s') \tag{6}$$

The policy is a mapping between the state space and the action space. In MDP, we usually define this mapping as a probability function $\pi(s, a)$ of taking action a when state is s. Given a policy and a particular state, the value function is defined as

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \tag{7}$$

where a discount factor γ discounts future rewards. Similarly, we define the value of taking action a in state s according to a policy as

$$Q^\pi(s, a) = E_\pi\{R_t | s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \tag{8}$$

An important issue in reinforcement learning is *behavioral variety*, which is the trade-off between *exploration* and *exploitation*. Exploration refers to the activity of evaluating the value of available actions. Exploitation means utilizing current knowledge of action values to maximize the return. The agent must decide to select the action that maximizes

the reward or the action that has not been fully explored. There are many techniques to balance the trade-off between exploration and exploitation. One of the simplest approaches is called the *ε-greedy* exploration method. In this method, instead of being greedy all the time, the agent takes non-greedy action once in a while, for example with the probability of ε. Another category of methods is called *softmax action selection* methods, among which the Gibbs or Boltzmann distribution is one of the most popular.

This method defines the rule of choosing an action with probability

$$P(s,a) = \frac{e^{Q(s,a)/\tau}}{\sum_{b=1}^{n} e^{Q(s,a)/\tau}} \quad (9)$$

where τ is a positive parameter called temperature. High temperatures cause all actions to be (nearly) equiprobable; low temperatures cause a greater difference in selection probability for actions that differ in their value estimates. When $\tau \to 0$, softmax selection becomes the greedy action selection. A more detailed introduction of reinforcement learning can be found in Sutton and Barto [21].

One of the most popular algorithms, which is also used in this study is the Q-learning algorithm introduced by Watkins and Dayan [22]. The key concept is that the action-value function $Q(s, a)$ is used directly to approximate the optimal value $Q^*(s, a)$. The update rule is defined as

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \quad (10)$$

where $0 \leq \alpha \leq 1$ stands for a learning rate, and $0 \leq \gamma \leq 1$ is the discount factor introduced previously. The $Q(s, a)$ values are usually stored in a lookup table, the so-called Q-table, and each entry represents an estimation of the Q-value of a state-action pair. A simple interpretation of the Q-learning algorithm is that the estimation of the optimal action-value function $Q(s, a)$ is a combination of past memory and new experience.

2.2.2. Simulated reinforcement learning

The proposed hybrid control scheme is based on simulated reinforcement learning as shown in Fig. 2. Instead of getting experience from the environment directly, the learning procedure of the controller is divided into two phases: a simulated learning phase and an implemented learning phase, as depicted in Fig. 4.

(1) *Simulated learning phase.* In this phase, the learning controller is trained by a simulator to learn as much as it can within a certain training period or training effort. This phase is a preliminary learning phase or a "guiding" procedure. The simulation model is not necessarily a perfect match of the environment, but it needs to provide the correct state-action information to guide the learning controller to get into the right "zone". Since the simulation training is carried out off-line, it takes only hours or days instead of hundreds or thousands of days in reality to make the learning controller find the optimal policy.

(2) *Implemented learning phase.* With some trained experience, the learning controller can be implemented into the actual environment. In the second learning phase, the learning controller is expected to learn and improve its performance further by directly communicating with the actual environment. This is considered a refined learning phase or a tuning procedure. No matter how good the training model is, there will always be deviations between the model and reality. During the second learning phase, the learning controller can correct the mistakes that may exist in the simulator, and discover experience that was not found in the simulation training.

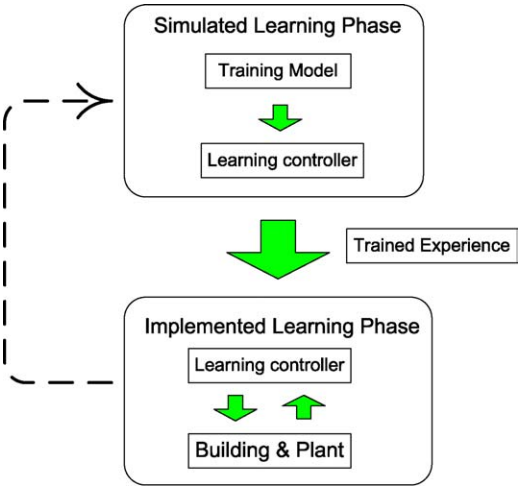

Fig. 4. Hybrid control scheme.

Simply said, in the hybrid control scheme, the simulated learning phase overcomes the slowness of pure reinforcement learning, and the implemented learning phase offers the opportunity to let the controller find the "true" optimum in the actual environment. The dashed line in Fig. 4 indicates that there is an opportunity to improve the performance of the simulated learning phase by calibrating the training model using the actual measurement during the implemented learning phase.

3. Overview of simulated and experimental control performance

The application of classic reinforcement learning to the optimal control of building active and passive thermal storage inventory has been investigated in simulation studies reported in [18,19,23]. In general, the results of these simulation studies confirm that reinforcement learning control is a feasible methodology to derive the optimal control policy for this specific problem. The learning

controller successively learns to pre-cool the building when the building is unoccupied, and charges or discharges the TES according to the utility rate structure. The operating cost savings do not reach the levels of model-based predictive optimal control, but are still substantial compared with conventional control strategies. Theoretically, the reinforcement learning algorithm can reach the true optimum given properly selected learning parameters and long enough learning time. The amount of required training is not realistic if the controller is directly implemented in a commercial building application. This constitutes the major drawback of the reinforcement learning control approach. This indicates that an Aristotelian learning control with no prior domain knowledge, i.e., a *tabula rasa*, is not going to be a practical solution and contextual information in some form needs to be introduced to expedite learning of the fundamental features of the problem, while reinforcement learning accommodates the fine-tuning of the controller. This inspires the generation of the hybrid learning control scheme introduced in this study. An experimental study had been carried out in the laboratory facility called Energy Resource Station at the Iowa Energy Center in Ankeny Iowa to validate the hybrid learning control approach and analyze its performance. In the companion paper [24], a detailed discussion of this experimental study will be presented.

References

[1] J.P. Conniff, Strategies for reducing peak air-conditioning loads by using heat storage in the building structure, ASHRAE Transactions 97 (1) (1991) 704–709.

[2] F.B. Morris, J.E. Braun, S.J. Treado, Experimental and simulated performance of optimal control of building thermal storage, ASHRAE Transactions 100 (1) (1994) 402–414.

[3] A. Rabl, L.K. Norford, Peak load reduction by preconditioning buildings at night, International Journal of Energy Research 15 (1991) 781–798.

[4] K.R. Keeney, J. Braun, A simplified method for determining optimal cooling control strategies for thermal storage in building mass, International Journal of HVAC&R Research 2 (1) (1996) 59–78.

[5] J.E. Braun, Reducing energy costs and peak electrical demand through optimal control of building thermal mass, ASHRAE Transactions 96 (2) (1990) 876–888.

[6] I. Andresen, M. Brandemuehl, Heat storage in building thermal mass: a parametric study, ASHRAE Transactions 98 (2) (1992) 910–918.

[7] J. Braun, Load control using building thermal mass, Journal of Solar Energy Engineering 125 (3) (2003) 292–301.

[8] G.P. Henze, M. Krarti, M.J. Brandemuehl, A simulation environment for the analysis of ice storage controls, International Journal of HVAC&R Research 3 (2) (1997) 128–148.

[9] G.P. Henze, M. Krarti, M. Brandemuehl, Guidelines for improved performance of ice storage systems, Energy and Buildings 35 (2) (2002) 111–127.

[10] G.P. Henze, R. Dodier, M. Krarti, Development of a predictive optimal controller for thermal energy storage systems, International Journal of HVAC&R Research 3 (3) (1997) 233–264.

[11] G.P. Henze, M. Krarti, The impact of forecasting uncertainty on the performance of a predictive optimal controller for thermal energy storage systems, ASHRAE Transactions 105 (2) (1999) 553–561.

[12] G.P. Henze, C. Felsmann, G. Knabe, Evaluation of optimal control for active and passive building thermal storage, International Journal of HVAC&R Research 9 (3) (2004) 259–275.

[13] G.P. Henze, D. Kalz, C. Felsmann, G. Knabe, Impact of forecasting accuracy on predictive optimal control of active and passive building thermal storage inventory, International Journal of HVAC&R Research 9 (3) (2003) 259–275.

[14] S. Liu, G.P. Henze, Impact of modeling accuracy on predictive optimal control of active and passive building thermal storage inventory, ASHRAE Transactions 110 (1) (2004) 4683.

[15] G.P. Henze, D. Kalz, S. Liu, C. Felsmann, Experimental analysis of model-based predictive optimal control for active and passive building thermal storage inventory, International Journal of HVAC&R Research 11 (2) (2005) 189–214.

[16] G.P. Henze, R. Dodier, Adaptive optimal control of a grid-independent photovoltaic system, Journal of Solar Energy Engineering 125 (1) (2003) 34–42.

[17] G.P. Henze, J. Schoenmann, Evaluation of reinforcement learning control for thermal energy storage systems, International Journal of HVAC&R Research 9 (3) (2003) 259–275.

[18] S. Liu, G.P. Henze, Reinforcement learning control for building active and passive thermal storage inventory, in: Proceedings of the SimBuild 2004, Boulder, CO, 2004.

[19] S. Liu, G.P. Henze, Reinforcement learning control for building active and passive thermal storage inventory, in: Proceedings of the 2005 International Solar Energy Conference, Orlando, FL, 2005.

[20] M.L. Littman, Algorithms for sequential decision making, Ph.D. Dissertation, Brown University, 1996.

[21] R.S. Sutton, A.G. Barto, Reinforcement learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[22] C. Watkins, P. Dayan, *Q*-learning, Machine Learning 8 (1992) 279–292.

[23] S. Liu, Analytical and experimental comparison of model-based, model-free, and hybrid learning control of active and passive building thermal storage inventory, Ph.D. Thesis, Department of Architectural Engineering, University of Nebraska-Lincoln, 2005.

[24] S. Liu, G.P. Henze, Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory. Part 2. Experimental results, Energy and Buildings, 38 (2006) 148–161.