NONLINEAR DYNAMICAL SYSTEMS APPROACH TO BUILDING ENERGY PREDICTION PROBLEMS

Y.Nakajima, M.Saito, J.Sugi and T.Matsumoto

Department of Electrical, Electronics and Computer Engineering
Waseda University
3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan
takashi@mse.waseda.ac.jp

ABSTRACT

Given a time series data, model dynamical systems are built using a hierarchical Bayesian scheme with feedforward neural nets and then the models are compared in terms of *marginal likelihood*. The model with the highest marginal likelihood is used for predictions. The algorithm is applied to building air-conditioning load prediction.

INTRODUCTION

When nonlinearity is present, time series prediction becomes a difficult task. The problem is particularly difficult when no functional form (equation) is known of the dynamics.

Given time series data, typically one constructs a model nonlinear dynamical system which fits to the given data, and makes predictions using the model. There are several important issues which need to be addressed:

(i) Which class of models should be used;
(ii) How should one estimate parameters associated with models without overfitting.

This paper models a non-autonomous nonlinear dynamical system by feedforward neural nets with a hierarchical Bayes scheme [1] and then applies the proposed algorithm to building air-conditioning load prediction problem.

Saving energy and reduction of CO_2 emissions are becoming critical for conservation of global and regional environments. The cost of electricity during night hours is typically much less than that of the daytime. Therefore, in electrically operated HVAC (Heating, Ventilation, and Air-Conditioning) systems, introduction of thermal energy storage systems can help level off electricity demand throughout the day and thus increase the over all operation efficiency of the power plants run by utility companies. Thermal energy storage systems therefore contribute to avoid construction of additional power plants and stability of power systems.

However, in reality thermal energy storage systems are often found not operating as efficiently as expected at the design stage. Typical reasons for this are: excessive storage of thermal energy leads to significant heat loss through tank surroundings; and peak hour operation of energy plants becomes necessary because stored energy is completely discharged early in a day. In order to overcome these problems, very good prediction algorithms are needed for predicting air-conditioning loads.

A prediction competition was organized by SHASE (Society of Heating, Air-conditioning, and Sanitary Engineers in Japan) [4] which we participated. The results are on those real data provided by the competition organizer.

Achilles heel of Bayesian method is its dependency on prior. Hierarchical Bayes considers a *family* of prior distributions parameterized by hyperparameters instead of a single prior. One can estimate hyperparameters given data and then one estimates the parameters in question. This way hierarchical Bayes approach enables algorithms less dependent on a particular prior. Hierarchical Bayes approach is also endowed with a natural structure for model comparisons which is extremely important.

FORMULATION

Problem :

Let data set $D := \left(\{x_t\}_{t=0}^N\ ,\ \{\boldsymbol{u}_t\}_{t=0}^N\right) \subset \mathbb{R} \times \mathbb{R}^m$ be given, where $\boldsymbol{u}_t$ is the input and x_t is the output. Given an additional input data $\{\boldsymbol{u}_t\}_{t=N+1}^T$, predict $\{x_t\}_{t=N+1}^T$.

Remark

In order to explain this formulation in terms of

building energy prediction problems, consider Table 1 which shows the data provided by [4] for prediction competition. The quantity to be predicted is thermal load which corresponds to x_t above. Other data including weather data and room environment data etc corresponds to $\boldsymbol{u}_t$ above.

Hypothesis $\mathcal{H}$

Hypothesis or model consists of the following:

(i) Architecture:
e.g., three-layer perceptron with h hidden units and a particular sigmoid function.

(ii) Likelihood:

$$P(\{x_t\}_{t=\tau}^{N} \mid \{\boldsymbol{u}_t\}_{t=\tau}^{N}, \boldsymbol{w}, \beta, \mathcal{H}) :=$$

$$\underbrace{\prod_{t=0}^{N-\tau} \frac{1}{Z_D(\beta)} \exp\left(-\beta E_D(x_{t+1} \mid x_t, x_{t-1}, \ldots, x_{t-\tau+1}; \boldsymbol{u}_t, \boldsymbol{w})\right)}_{\text{noisy dynamics}}$$

$$\times \quad \underbrace{P(x_{\tau-1}, \ldots, x_0 \mid \mathcal{H})}_{\text{initial state uncertainty}} \tag{1}$$

$$E_D(x_{t+1} \mid x_t, x_{t-1}, \ldots, x_{t-\tau+1}; \boldsymbol{u}_t, \boldsymbol{w}) := \frac{1}{2}(x_{t+1} - f(x_t, x_{t-1}, \ldots, x_{t-\tau+1}; \boldsymbol{u}_t, \boldsymbol{w}))^2 \tag{2}$$

where $f(\cdot)$ is neural net output, $\boldsymbol{w} \in \mathbb{R}^k$ the weight parameters of a particular architecture, β (unknown) uncertainty level, and $Z_D(\beta)$ the normalization constant, and τ is (unknown) order of dynamics.

(iii) Prior for $\boldsymbol{w}$:

$$P(\boldsymbol{w} \mid \boldsymbol{\alpha}, \mathcal{H}) := \prod_{c=1}^{C} \frac{1}{Z_W(\alpha_c)} \exp(-\alpha_c E_{W_c}(\boldsymbol{w}_c)) \tag{3}$$

where $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ are decomposed into groups:

$$\boldsymbol{w} := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C) \ , \ \boldsymbol{w}_c \in \mathbb{R}^{k_c}, \tag{4}$$

$$\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_C) \ , \ \alpha_c \in \mathbb{R} \tag{5}$$

$\exp(-\alpha_c E_{W_c}(\boldsymbol{w}_c))/Z_W(\alpha_c)$ represents the prior belief on how $\boldsymbol{w}_c$ should be distributed with (unknown) α_c and $Z_W(\alpha_c)$ is the normalization constant.

(iv) Prior for $(\boldsymbol{\alpha}, \beta)$, hyperparameters: $P(\boldsymbol{\alpha}, \beta \mid \mathcal{H})$

(v) Prior for $\mathcal{H}$: $P(\mathcal{H})$

It is important to note that (1) and (2) represent a *non-autonomous* dynamical system because of the presence of $\boldsymbol{u}_t$. Decomposition (4) of weight parameters and associated decomposition (5) of hyperparameters are also important. Typically, a subvector $\boldsymbol{w}_c$ consists of those weights between each input variable to feedforward neural net and hidden units so that dim $\boldsymbol{w}_c = h$, the number of hidden units. Another typical $\boldsymbol{w}_c$ consists of the biases for hidden units, and finally the bias for output unit. Note that the number of inputs is *not* dim $\boldsymbol{u}_t$, rather it is dim $\boldsymbol{u}_t + \tau$, where τ is the order of dynamical system (see Fig. 1).

Note also that (3) is not a smoothness penalty. Rather, it represents how $\boldsymbol{w}_c$ is concentrated around the origin, where α_c represents the degree of concentration. Therefore, if a particular α_c is very large compared with others, then $\boldsymbol{w}_c$ is highly concentrated around the origin and $\boldsymbol{w}_c$ may be unnecessary. Therefore, the Hierarchical Bayesian algorithm is naturally endowed with the ability of distinguishing relevant data from irrelevant data.

The goal of the prediction problem is to compute the predictive distribution $P(\{x_t\}_{t=N+1}^{T} \mid \{\boldsymbol{u}_t\}_{t=N+1}^{T}, D)$ under (i) – (v). This paper first computes three levels of posterior distributions as shown in Figure 1 and use them to compute the predictive distribution.

Level 1

$$P(\boldsymbol{w} \mid D, \boldsymbol{\alpha}, \beta, \mathcal{H}) = \frac{P(D \mid \boldsymbol{w}, \beta, \mathcal{H}) \ P(\boldsymbol{w} \mid \boldsymbol{\alpha}, \mathcal{H})}{P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H})}$$

Level 2

$$P(\boldsymbol{\alpha}, \beta \mid D, \mathcal{H}) = \frac{P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H}) \ P(\boldsymbol{\alpha}, \beta \mid \mathcal{H})}{P(D \mid \mathcal{H})}$$

Level 3

$$P(\mathcal{H} \mid D) = \frac{P(D \mid \mathcal{H}) \ P(\mathcal{H})}{P(D)}$$

$$\Downarrow$$

Predictive Distribution

$$P(\{x_t\}_{N+1}^{T} \mid \{\boldsymbol{u}_t\}_{N+1}^{T}, D) = \sum_{\mathcal{H}} \int\int\int P(\{x_t\}_{N+1}^{T} \mid \{\boldsymbol{u}_t\}_{N+1}^{T}, \boldsymbol{w}, \beta, \mathcal{H}) \times P(\boldsymbol{w}, \boldsymbol{\alpha}, \beta, \mathcal{H} \mid D) d\boldsymbol{w} d\boldsymbol{\alpha} d\beta$$

Figure 1: Hierarchical Bayesian Structure

PREDICTION

Fact 1 (Level 1: Posterior for $\boldsymbol{w}$)

The posterior of $\boldsymbol{w}$ given $(D, \boldsymbol{\alpha}, \beta, \mathcal{H})$ is

$$P(\boldsymbol{w} \mid D, \boldsymbol{\alpha}, \beta, \mathcal{H}) = \frac{\frac{\exp(-M(\boldsymbol{w};\boldsymbol{\alpha},\beta))}{Z_D(\beta)Z_W(\boldsymbol{\alpha})}}{P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H})} \quad (6)$$

where $M(\cdot)$ is the merit function,

$$M(\boldsymbol{w};\boldsymbol{\alpha},\beta) := \beta E_D(\boldsymbol{w}) + \sum_{c=1}^{C} \alpha_c E_{W_c}(\boldsymbol{w}_c) \quad (7)$$

and hence the most probable $\boldsymbol{w}$, called $\boldsymbol{w}_{\rm MP}$, is given by

$$\boldsymbol{w}_{\rm MP} = \arg\min_{\boldsymbol{w}} M(\boldsymbol{w};\boldsymbol{\alpha},\beta) \ . \quad (8)$$

Fact 2 (Level 2: Posterior for $(\boldsymbol{\alpha}, \beta)$)

If $P(\boldsymbol{\alpha}, \beta \mid \mathcal{H})$ is log flat, then the most probable hyperparameters are given by

$$(\boldsymbol{\alpha}_{\rm MP}, \beta_{\rm MP}) = \arg\max_{\boldsymbol{\alpha},\beta} P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H}) \quad (9)$$

so that the following gradient information can be used for finding $(\boldsymbol{\alpha}_{\rm MP}, \beta_{\rm MP})$:

$$\frac{\partial}{\partial\beta}\log P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H}) \approx -E_D(\boldsymbol{w}_{\rm MP}) - \frac{1}{2}{\rm Tr}\boldsymbol{A}^{-1}\boldsymbol{B}_D - \frac{\partial}{\partial\beta}\log Z_D(\beta) \quad (10)$$

where $\boldsymbol{A}$ is the Hessian of $M(\cdot)$ evaluated at $\boldsymbol{w}_{\rm MP}$, Tr stands for a trace of a matrix, E_D is defined by (2) and $\boldsymbol{B}_D$ is the Hessian of E_D evaluated at $\boldsymbol{w}_{\rm MP}$,

$$\frac{\partial}{\partial\alpha_c}\log P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H}) \approx -E_{W_c}(\boldsymbol{w}_{c\,\rm MP}) - \frac{\partial}{\partial\alpha_c}\log Z_W(\boldsymbol{\alpha}) - \frac{1}{2}{\rm Tr}\boldsymbol{A}^{-1}\boldsymbol{B}_c \quad (11)$$

where $\boldsymbol{B}_c$ is the Hessian of E_{W_c} evaluated at $\boldsymbol{w}_{\rm MP}$.

Fact 3 (Level 3: Posterior for $\mathcal{H}$ (model comparison))

If $P(\mathcal{H})$ is flat, then the most probable model is given by

$$\mathcal{H}_{\rm MP} = \arg\max_{\mathcal{H}} P(D \mid \mathcal{H}) \quad (12)$$

Fact 4 (Predictive distribution)

$$P(\{x_t\}_{N+1}^T \mid \{\boldsymbol{u}_t\}_{N+1}^T, D) = \sum_{\mathcal{H}} \int\int\int P(\{x_t\}_{N+1}^T \mid \{\boldsymbol{u}_t\}_{N+1}^T, \boldsymbol{w}, \beta, \mathcal{H}) \times P(\boldsymbol{w}, \boldsymbol{\alpha}, \beta, \mathcal{H} \mid D) d\boldsymbol{w} d\boldsymbol{\alpha} d\beta \quad (13)$$

If

$$P(\{x_t\}_{N+1}^T \mid \{\boldsymbol{u}_t\}_{N+1}^T, \boldsymbol{w}, \beta_{\rm MP}, \mathcal{H}_{\rm MP}) \approx \prod_t \frac{1}{Z_D(\beta_{\rm MP})} \times \exp\left\{-\frac{\beta_{\rm MP}}{2}\left(x_{t+1} - f(x_t, \ldots, x_{t-\tau+1}; \boldsymbol{u}_t, \boldsymbol{w}_{\rm MP}) - \frac{\partial f}{\partial \boldsymbol{w}}^T (\boldsymbol{w} - \boldsymbol{w}_{\rm MP})\right)^2\right\} \quad (14)$$

$$P(\boldsymbol{w} \mid D, \alpha, \beta, \mathcal{H}) \approx \frac{1}{(2\pi)^{h/2}\det \boldsymbol{A}^{-1/2}} \times \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\rm MP})^T \boldsymbol{A} (\boldsymbol{w} - \boldsymbol{w}_{\rm MP})\right) \quad (15)$$

then the most probable prediction, $x_{t,\rm MP}$ is given by

$$x_{t+1,\rm MP} = f(x_{t,\rm MP}, \boldsymbol{u}_t, \boldsymbol{w}_{\rm MP}) \ , \ N \le t \le T-1 \ . \quad (16)$$

Log marginal likelihood $-2\log P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H})$ is sometimes called ABIC[2] or evidence for hyperparameters[3], and marginal likelihood at the next hierarchy $P(D \mid \mathcal{H})$ is sometimes called evidence for model[3]. The quantity proposed in [2], $-2\log P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H}) + 2\dim(\boldsymbol{\alpha}, \beta)$ is different from $-2\log P(D \mid \mathcal{H})$, however.

AIR-CONDITIONING LOAD PREDICTION

Problem Description

"The First International Benchmark Test of Air-conditioning Load Prediction Methods for Optimum Operation of Thermal Energy Storage Systems" was organized by The Technical Committee for Optimization of Thermal Energy Storage Systems (TC-OTES), a division of The Society of Heating, Air-conditioning, and Sanitary Engineers of Japan (SHASE) [4]. The building from which the data are taken is eleven storied with total area 28,481 m^2, which is large. Most rooms are used for research purposes. The data provided consist of hourly values of the quantities shown in Table 1 for each of the fourth floor through tenth floor between June 1 and July 31, 1996. Given additional weather data for August 1 $\sim$ 31, contestants are asked to predict the *total sum* (fourth

floor through tenth floor) of thermal load instead of the load of individual floors.

Architecture for Nonlinear Dynamics

First note that a feedforward neural net is generally described by

$$y = \sum_{i=1}^{h} a_i \cdot \sigma \left(\sum_{j=1}^{n} w_{ij} z_j + \theta_i \right) + \eta$$

where $z_j, j = 1, ..., n$, are inputs, y is output, $(w_{ij}, a_i, \theta_i, \eta)$ represent weight parameters and $\sigma(\cdot)$ is so called sigmoid function. In the building energy prediction problem discussed here, y represents x_{t+1}, the target value, a part of z_j represents weather data and room environment data, and the remaining z_j are (x_t, x_{t-1}), the target value at time t and $t-1$. As was described in FORMULATION, training data set D consists of $(\{x_t\}_{t=0}^N, \{\boldsymbol{u}_t\}_{t=0}^N)$ so that neural net weight parameters are adjusted to fit to D. In any prediction problem, however, prevention of *overfitting* is critical. Namely, while it is possible to fit to the training data set D in an extremely accurate manner, such a prediction algorithm would overfit to D and will have poor prediction capabilities. In order to avoid overfitting, one often considers

$$\frac{\beta}{2} \parallel x_{t+1} - f(x_t, \boldsymbol{u}_t, \boldsymbol{w}) \parallel^2 + \frac{\alpha}{2} \parallel \boldsymbol{w} \parallel^2 \qquad (17)$$

where $f(\cdot)$ is neural net output so that the first term represents the data fitting term. The second term is the penalty on the weight parameter, where $\boldsymbol{w}$ stands for $\{z_{ij}\}, \{a_i\}, \{\theta_i\}, \eta$. Equation (17) contains two extra parameters α and β which control how much emphasis should be put on the data fit term and penalty term. When $\beta \gg \alpha$, overfitting would occur, while $\beta \ll \alpha$, underfitting would take place. α and β are called *hyperparameters* and should be estimated from the data set D. It should be noted that weight $\boldsymbol{w}$, in this paper, is decomposed as described in FORMULATION in such a way that relevant of each input ca be determined.

Of the six possible input data (four weather data and two room environment data), only three of them were used for prediction purposes; OT (outside temperature), RT (room temperature) and RH (room humidity). This is a decision based on several preliminary data analyses and our earlier experience with ASHRAE competition [5]. During our preliminary data analysis phase, we found that the target value x_t, the total load at time t and x_{t-1}, the load at $t-1$, have effects on x_{t+1}. Figure 2 shows our architecture where $u_{\mathrm{t,\ time}}$ stands for time variable, $u_{\mathrm{t,\ OT}}$, $u_{\mathrm{t,\ RT}}$ and $u_{\mathrm{t,\ RH}}$ indicate OT, RT, and RH at time t. This is for workdays. We found it extremely difficult to train neural nets for weekends and holidays. At least part of this difficulty appears to be attributable to the fact that the number of data for weekends and holidays are much smaller than that of workdays. A straightforward way of reducing the number of parameters is to reduce the number of inputs to neural nets. For this reason, x_t and x_{t-1} were eliminated for weekends and holidays, which would decrease prediction capabilities.

Training

With the architecture described in the previous section, training was performed by the algorithm given in FORMULATION. Model selection (model comparison) described in Fact 3 amounts to select a neural net with the optimal nubmer of hidden units. Theoretically, this is given by (12) which is extremely difficult to compute. Under the assumption that $P(D \mid \boldsymbol{\alpha}, \beta, \mathcal{H})$ has a reasonally sharp peak at $(\boldsymbol{\alpha}_{\mathrm{MP}}, \beta_{\mathrm{MP}})$ (see (9)) one can assume

$$\mathcal{H}_{\mathrm{MP}} = \arg\max_h P(D \mid \boldsymbol{\alpha}_{\mathrm{MP}}, \beta_{\mathrm{MP}}, \mathcal{H})$$

This is our scheme for selecting best h. Figure 3 shows plots of

$$\log P(D|\boldsymbol{\alpha}_{\mathrm{MP}}, \beta_{\mathrm{MP}}, \mathcal{H}) \qquad (18)$$

against the number of hidden units h, between 1 and 9 for workdays. Multiple plots correspond to different initial conditions for the optimization of $M(\boldsymbol{w}; \boldsymbol{\alpha}, \beta)$ with respect to $\boldsymbol{w}$ (see (8)). Log marginal likelihood in the figure stands for (18). The larger the log marginal likelihood, the better the model.

Prediction

The model $\mathcal{H}_{\mathrm{MP}}$ with the highest value of (18) was chosen ($h = 5$). Now we have chosen the best model $\mathcal{H}_{\mathrm{MP}}$, best parameters $\boldsymbol{w}_{\mathrm{MP}}$ and best hyperparameters $(\boldsymbol{\alpha}_{\mathrm{MP}}, \beta_{\mathrm{MP}})$. Our prediction is done by (16), i.e., the predicted value is given by the dynamical system described by this equation. Figure 4 (a) and (b) compare our predictions (solid line) with the actual data which was disclosed after the competition was over. The predictions appear to be excellent.

CONCLUSION

A hierarchical Bayesian algorithm was used to train feedforward neural nets to construct nonlinear dynamical systems which represent given time series data. The algorithm was applied to make predictions of thermal loads of air-condition coils in a building. The results look encouraging. Due to the fact that time was very short in preparing

the predictions, several issues were not thoroughly discussed;

(i) More detailed comparisons of $P(D|\mathcal{H})$ should have been made with different τ, the order of the dynamical system (see (1));
(ii) Predictions for weekends and holidays need improvements.

In this competition, various weather data were available on the hourly basis. This may not be always possible in more realistic situations. "The Second International Benchmark Test of Air-conditioning Load Prediction Method for Optimum Operation of Thermal Energy Storage Systems", the second phase of [4] is under way for such environments.

ACKNOWLEDGMENT

Measured data in benchmark test could be available by courtesy of Tokyo Electric Power Company.

REFERENCES

[1] Matsumoto, T., Hamagishi, H. and Y. Chonan [1997], "A Hierarchical Bayes Approach to Nonlinear Time Series Prediction with Neural Nets", *Proceedings of the 1997 International Conference on Neural Networks*, 2028-2033.

[2] Akaike, H. [1980], "Likelihood and the Bayes procedure", In *Bayesian Statistics*, J.M.Bernardo, M.H. DeGroot, D.V.Lindley and A.F.M.Smith, eds, University Press, Valencia, Spain, 143-166.

[3] MacKay, D.J.C. [1991], "Bayesian Methods for Adaptive Models, PhD thesis", California Inst. Tech. Pasadena 1991.

[4] Society of Heating, Air-conditioning and Sanitary Engineers in Japan (SHASE) [1997], " International Benchmark Test of Air-conditioning Load Prediction Methods for Optimum Operation of Thermal Energy Storage Systems", *http://www.t3.rim.or.jp/˜bmtest.*

[5] Haberl,J.S. S.Thamilseran [1996], "The Great Energy Predictor Shootout II : Measuring Retrofit Savings", *ASHRAE Transactions* 1996 Vol.102 PART 2.

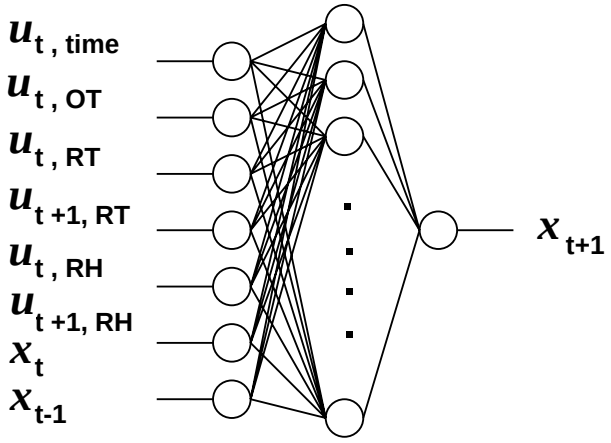
[6] Y.Nakajima, J.Sugi, M.Saito, H.Hamagishi, D.Hattori and T.Matsumoto [1998], "Hierarchical Bayesian Neural Nets for Air-conditioning Load Prediction : Nonlinear Dynamics Approach", *1998 IEEE World Congress on Computational Intelligence* 1948-1953.

[7]T.Matsumoto, Y.Nakajima, H.Hamagishi, J.Sugi and M.Saito [1998], "From Data to Nonlinear Dynamics : A Hierarchical Bayes Approach with Neural Nets", *IEEE Workshop on Neural Networks for Signal Processing* VIII, 333-342.

[8]Y.Nakajima, J.Sugi, M.Saito, H.Hamagishi and T.Matsumoto [1998], "A Hierarchical Bayes Algorithm for Air-conditioning Load Prediction : Nonlinear Dynamics Approach", *The Fifth International Conference on Neural Infomation Processing*, 1347-1350.

Table 1: Training data

data		unit
weather data	outside air temperature	C
	outside air humidity	%
	overall horizontal solar radiation	W/m^2
	wind velocity	m/s
room environment data	average room temperature of the n-th floor	C
	average room humidity of the n-th floor	%
system operation flag	air-conditioning systems operation flag	-
coil thermal load	total thermal load of the n-th floor air-conditioning coil	kcal/h



OT: outside temperature, RT: room temperature, RH: room humidity,
x_t: thermal load of air-conditioning coil

Figure 2: Architecture of nonlinear dynamical system for predictions

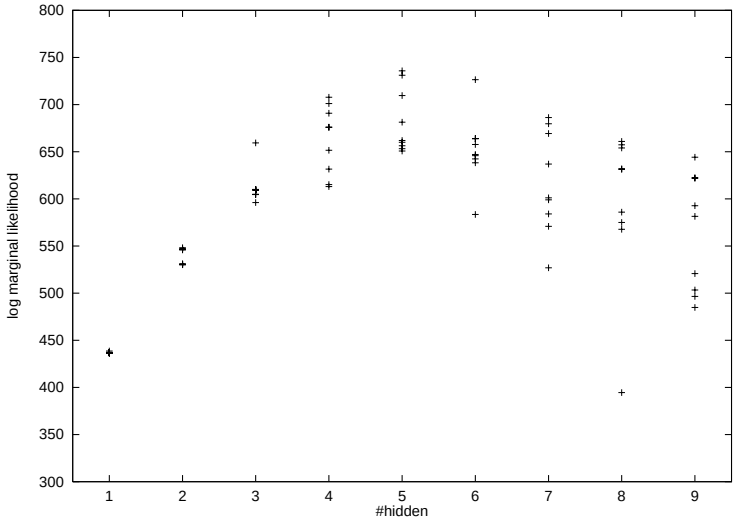
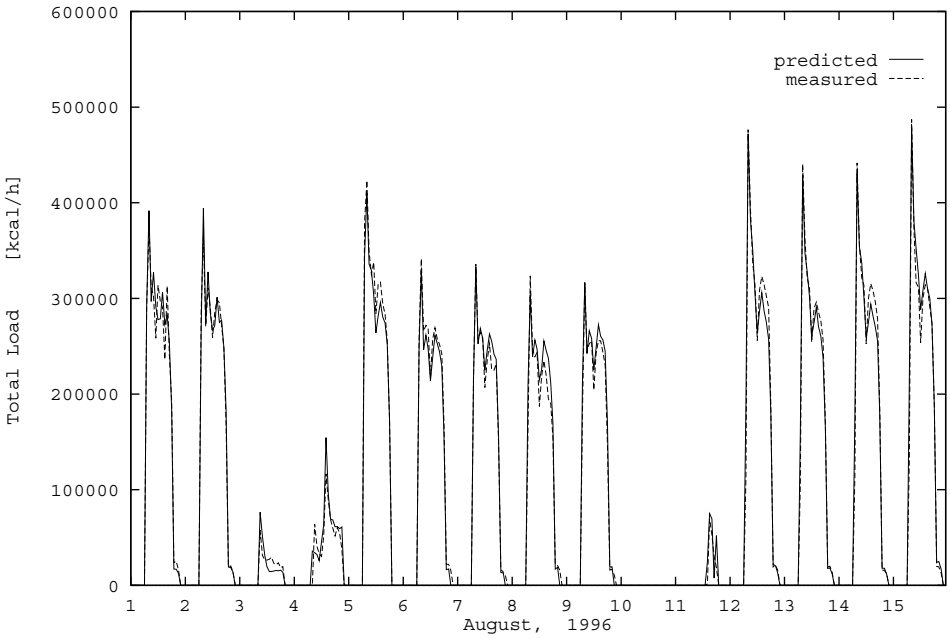
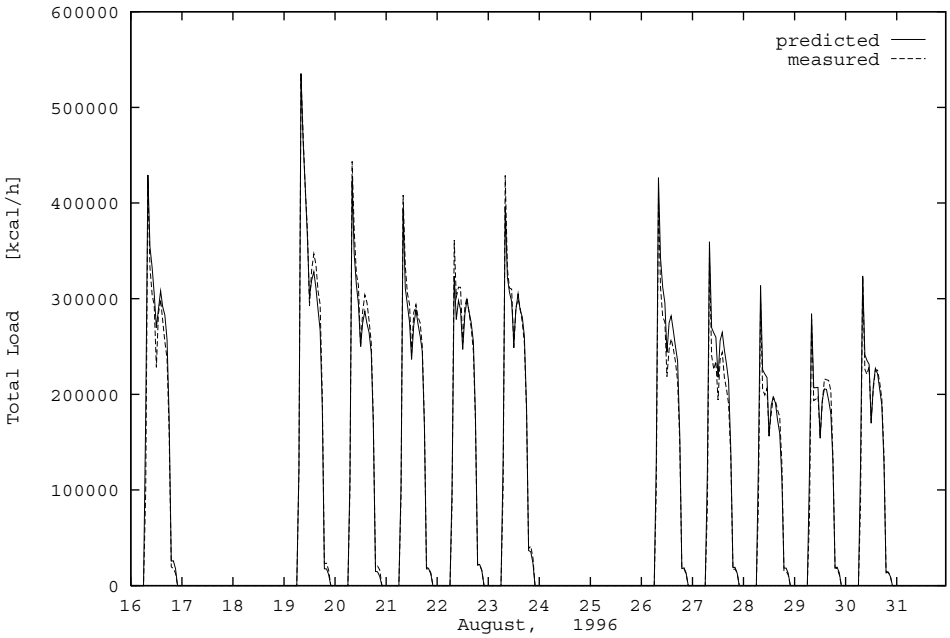


Figure 3: $\log P(D \mid \alpha_{\mathrm{MP}}, \beta_{\mathrm{MP}}, \mathcal{H})$ vs. h



(a) From August 1 to August 15



(b) From August 16 to August 31

Figure 4: Our predictions compared with measured data which was disclosed after the competition.