ESL-TR-94/12-07

November 1994. To be published in the 1995 ASME Int'l Solar Energy Conference Proceedings.

BIAS IN PREDICTING ANNUAL ENERGY USE IN COMMERCIAL BUILDINGS WITH REGRESSION MODELS DEVELOPED FROM SHORT DATA SETS

S. Katipamula, T.A. Reddy*, D.E. Claridge*
Pacific Northwest Laboratory, K5-20, P.O Box 999
Richland, Washington 99352

ABSTRACT

An empirical or regression modeling approach is simple to develop and easy to use compared to use of detailed hourly simulations. Therefore, regression analysis has become a widely used tool in the determination of annual energy savings accruing from energy conserving retrofits. The regression modeling approach is accurate and reliable if several months of data (more than six months) are used to develop the model. If such is not the case, the regression models can, unfortunately, lead to significant errors in the prediction of the annual energy consumption.

Issues relating to bias in regression models identified from short data sets are discussed in this paper. First, the physical reasons for the differences between the predictions of the annual energy consumption based on a short data set model and on a long data set model are discussed. Then, the errors associated with the multiple linear regression model are evaluated when applied to short data sets of monitored data from large commercial buildings in Texas.

The analysis shows that the seasonal variation of the outdoor dry-bulb and dew-point temperature causes significant errors in the models developed from short data sets. The MBE (mean bias error) from models based on short data sets (one month) varied from +40% to -15%, which is significant. Hence, due care must be exercised when applying the regression modeling approach in such cases.

*Dept. of Mechanical Engineering, Texas A&M University, College Station, TX 77843-3123.

INTRODUCTION

The performance of energy conservation retrofits is being increasingly assessed by direct measurement of energy savings. (Fels 1986, MacDonald and Wasserman, 1988, Kissock et al., 1992 and Reddy et al., 1994a). Commonly used methods to evaluate retrofit performance include: (i) direct utility bill comparison from pre- to post-retrofit energy use, (ii) use of pre-retrofit empirical energy consumption model with post-retrofit conditions (Kissock et al. 1992 and Katipamula et al., 1994a), (iii) use of simplified loads and systems models (Katipamula and Claridge, 1993), and (iv) detailed hourly simulation (DOE2.1, BLAST, etc.) of pre- and post-retrofit energy consumption (Kaplan et.al., 1990).

The simplest amongst the savings measurement methods is direct utility bill comparison, which involves comparing utility bills for each month prior to retrofit to those after the retrofit. However, it is limited to simple non-weather dependent retrofits (such as lighting). On the other hand, the detailed hourly simulation method is probably the most accurate, but requires advanced skills and several critical inputs; thus it is time consuming. Empirical or regression models, based on engineering principles (relating to building description, HVAC equipment, occupancy and internal load, etc.) are simple to develop and easy to use as compared to the detailed hourly simulation models. Therefore, regression analysis has become a useful tool in the measurement of energy savings from energy conserving retrofits.

Residential energy consumption (heating and cooling related) is usually a strong function of the outdoor dry-bulb temperature (Fels, 1986), whereas the energy consumption of a large commercial building is a complex function of climatic conditions (dry-bulb and dew-point temperatures, and solar loads), building characteristics (loss coefficients, heat capacity, internal loads), building usage (12-hour, or 24-hour, amount of fresh air intake ...), system characteristics (total air flow, hot and cold deck supply temperatures, economizer cycle), and type of Heating, Ventilation, and Air Conditioning (HVAC) equipment used. Therefore, Multiple Linear Regression (MLR) models are well suited to model the energy consumption of large commercial buildings (Reddy et al., 1994b and Katipamula et al., 1994a).

It has been suggested that the regression modeling approach is reliable if several months (six months or more) of data are used for model identification (Kissock et al., 1993). Unfortunately, if fewer than six months of daily monitored data are available, the regression models may lead to significant errors in the prediction of the annual energy consumption. Rachlin et al. (1986) studied how the parameter estimates of the model identified using the PRInceton Scorekeeping Method (PRISM, Fels 1986) varied for different estimation periods. They concluded that twelve consecutive monthly readings are required for robust estimation of normalized annual energy consumption using utility billing data. Kissock et al. (1993) showed that the annual energy use in a commercial building in Central Texas with dual-duct constant volume system (DDCV) can be significantly over-or under-estimated depending on the size and properties of the data sets used. The study showed that for the buildings considered, the maximum deviation of annual cooling energy use was about 20% when a one month period was used with an average deviation of 4% depending on the size of the data set. The maximum deviation of annual heating energy use was as much as 400% with an average around 20%. They also suggested that models identified from short data sets which included the spring and the fall period had less bias than those identified from the winter and the summer period. All the comparisons were based on daily energy use models.

This study is an extension of the previous work by Kissock et al. (1993). First, physical reasons for the differences between the predictions of the annual energy consumption based on a short data set model and the predictions based on a long data set model are highlighted. Second, regression models are developed using engineering principles and applied to short data sets for several different buildings that have the two most commonly used HVAC systems (dual-duct constant air volume [DDCV] and dual-duct variable air volume [VAV] systems) in Central Texas. Finally, an analysis of how the MBE (mean bias error) varies as a function of outdoor dry-bulb temperature for one-month data sets is presented.

BACKGROUND

In this section, we highlight the physical reasons for the differences between the annual energy consumption predicted from a model based on a short data set and the annual energy consumption predicted from a model based on a long data set (in this analysis a long data set consists of an entire year data) are highlighted. Four commonly used regression models to model the building energy consumption (both for residential and non-residential buildings) are the: (i) 2P, (ii) 3P, (iii) 4P and (iv) MLR models (Fels, 1986, Kissock, 1992, Ruch, et al., 1992 and Katipamula et al., 1994a). While 2P, 3P, and 4P models are generally applicable to residential and small commercial buildings (single zone type buildings), the use of MLR model is generally more appropriate for large (multi-zone type buildings) commercial buildings (Ruch et. al., 1992, and Katipamula et al., 1994a).

Since this study is focused on large commercial buildings, only the MLR model is analyzed. The analysis can be easily extended to other model types as well. Unlike heating energy consumption, the cooling energy consumption has two components, sensible and latent. In most residential and small commercial buildings the latent cooling is a small fraction of the sensible cooling consumption. Therefore, the total cooling energy use is assumed to have the same functional form as the sensible cooling energy consumption. However, this assumption is not valid for large commercial building where latent cooling can be a significant portion of the total cooling (depending on the fraction of outdoor air intake, Reddy et al., 1994a and Katipamula et al., 1994a). The 2P, 3P, and 4P models fall in the former category, where the total cooling energy consumption is assumed to have the same functional form as the sensible cooling energy consumption.

A previous study by Katipamula et al., (1994b), found that daily regression models (i.e. models developed from daily energy use data) were more appropriate than monthly, hourly or hour of the day regression models for estimating the energy savings from energy conserving retrofits. Therefore, only daily models are discussed in this study.

Previous studies have suggested that a piecewise MLR model explains more variation while modeling/predicting the cooling energy use ($\dot{E}_c$) in large commercial buildings with either a DDCV or a VAV system (Reddy et al., 1994a and Katipamula et al., 1994a). Also, the $\dot{E}_c$ in large commercial buildings, operating 24 hours a day, may show a segmented linear behavior as a function of outdoor dry-bulb temperature (T_o) (Kissock et al., 1992 and Katipamula et al., 1994). One of the reasons for the segmented linear behavior is due to the hot deck reset schedule in a DDCV system. In a VAV system, the change in slope as a function of T_o occurs when the system reaches its minimum air flow condition (In a VAV system air flow is modulated to meet the zone load, but this modulation stops after a certain minimum. Because the fan motors can not be modulated below 30% of the full load rated power).

In addition to T_o, the other variables which effect $\dot{E}_c$ are: outdoor dew-point temperature (T_{dp}), solar radiation (q_{sol}) and internal gains from lights and equipment (q_i). In most large commercial buildings the major portion of the latent cooling load is from ventilation, which is a strong function of outdoor dew-point (T_{dp}). However, while performing regression analysis, a better way to handle the latent contribution is to use $(T_{dp} - T_s)^+$, where T_s is the mean surface temperature of the cooling coil and "+" indicates that the term is set equal to zero when $T_{dp} < T_s$.

Since $\dot{E}_c$ is a function of several variables, MLR models are more appropriate than the single variable models. In regression modeling a change in slope can be handled by introducing an indicator variable. The value of the indicator variable, I, is set equal to 1 for T_o values to the right of the change point, indicating the presence of the change point. For the rest of the data (T_o values to the left of the change point) the indicator is set equal to zero (Daniel et al., 1980). Thus the regression model for both DDCV and VAV systems assumes the following functional form (Katipamula et. al., 1994a):

$$\dot{E}_c = \alpha + \beta_0 T_o + \beta_1 I + \beta_2 I T_o + \beta_3 T_{dp}^+ + \beta_4 q_i + \beta_5 q_{sol} \quad (1)$$

where α, β_0, β_1, β_2, β_3, β_4, and β_5 are regression coefficients. The coefficient β_1 represents the offset of the indicated observations (T_o values right of the change point) from the value α. The coefficient β_2 indicates the extent to which the right hand slope is larger than the slope to the left. If the change point does not exist in the data, then the coefficients β_1 and β_2 will not be statistically significant (i.e., both β_1 and β_2 will be near zero). The annual energy use (AEU) for the above model is given by:

$$\begin{aligned} AEU_{MLR} = & 365(\alpha + \beta_0 \overline{T_o} + \beta_3 \overline{T_{dp}^+} + \beta_4 \overline{q_i} + \beta_5 \overline{q_{sol}}) \\ & + (365 - n_1)(\beta_1 + \beta_2 \overline{T_r}) \end{aligned} \quad (2)$$

where, n_1, is the number of data points to the right of the change point. The fraction error (FE) in predicting AEU, with such model, is given by:

$$FE_{MLR} = 1 - \frac{365(\alpha' + \beta_0' \overline{T_o} + \beta_3' \overline{T_{dp}^+} + \beta_4' \overline{q_i} + \beta_5' \overline{q_{sol}}) + (365 - n_1')(\beta_1' + \beta_2' \overline{T_r}')}{365(\alpha + \beta_0 \overline{T_o} + \beta_3 \overline{T_{dp}^+} + \beta_4 \overline{q_i} + \beta_5 \overline{q_{sol}}) + (365 - n_1)(\beta_1 + \beta_2 \overline{T_r})} \quad (3)$$

where all estimates with "primes" pertain to the short data set and FE is defined as:

$$FE = \frac{(AEU_{long} - AEU_{short})}{AEU_{long}}$$

We note from the expression for the fractional error, that we need to know not only the regression parameter estimates based on a short data set but also those based on a long data set. There is no hard and fast rule to determine a priori the regression parameters based on a long data set for all building classes. Since the building energy consumption, in large commercial buildings, is a complex function of several variables, the regression parameters vary widely from building to building (Kissock et. al., 1992). Any generalization to quantify the fractional error has to be made by studying several buildings in each class (Office, Institutional, Hospital, Library, etc.). Therefore, in this study an alternate approach is used to quantify the errors (quantifying the mean bias error of short data sets as a function of regressor variables).

Before we address the source of errors associated with short data set models, the affect of collinearity between regressor variables is addressed. The MLR analysis assumes the regressor variables are independent of each other. Multicollinearity between the regressor variables results in large uncertainty bounds for the regression coefficients leading to model uncertainty. Therefore, in the next section the collinearity between regressor variables for a short data set is analyzed.

COLLINEARITY BETWEEN VARIABLES

Collinearity between regressor variables is a potential problem with data sets. It occurs whenever one (or more) of the regressor variables is a linear function of one (or more) other regressor variable. When regressor variables are related to each other, the estimates from the model can be misleading; therefore, it is important to understand the relationship of the regressor variables with one another. There are several tests to measure the degree of collinearity between independent variables. A common, though not completely adequate, measure of the degree of collinearity between two independent variables is the square of the sample correlation r_{12}^2 where subscript 12 refers to variable 1 against 2 (Weisberg, 1985). Exact collinearity corresponds to $r_{12}^2 = 1$; noncollinearity corresponds to $r_{12}^2 = 0$.

Another rule of thumb (Mullet, 1976) is that when the variance inflation factor ($[1 - r_{12}]^{-1}$) is less than 10, collinearity between independent variable 1 and 2 is considered insignificant. A rule of thumb suggested by Draper and Smith (1981) is that if the simple correlation between two variables is larger than the correlation of one or either variable with a dependent variable then collinearity effects can be important.

To study the effect of collinearity we selected daily cooling energy use data of a large institutional building in Central Texas with a DDCV system. A short data set is assumed to consist of one month's daily data only while a long data set is assumed to have data over an entire calendar year. The ratio of the correlation between regressor variables to the correlation between the regressor variable and the dependent variable for one-month data sets and long data set have been computed from the data (Figure 1). The rule suggested by Draper and Smith (1981) suggests that the collinearity effects are important when the ratio is greater than one. Figure 1 (a) shows the ratio ($R^2_{T_oT^+_{dp}}/R^2_{T_oE_c}$) and ($R^2_{T_oq_i}/R^2_{T_oE_c}$) for one-month data sets and for the year long data set. For the long data set, the collinearity between T_o and q_i is small, but there is some collinearity between T_o and T^+_{dp}. The ratio ($R^2_{T_oT^+_{dp}}/R^2_{T_oE_c}$) is greater than 1 for several one-month data sets (April, May and August), while the ratio $R^2_{T_oq_i}/R^2_{T_oE_c}$ is always less than 1. T_o accounts for over 60% of the variance in $\dot{E}_c$ (plot with triangle symbols), with the August data set being an exception.

Figure 1 (b) shows the variation of the ratios ($R^2_{T^+_{dp}T_o}/R^2_{T^+_{dp}E_c}$) and ($R^2_{T^+_{dp}q_i}/R^2_{T^+_{dp}E_c}$) for one-month data sets and for the year long data set. For the year long data set, there is some collinearity between T^+_{dp} and T_o. For the monthly data sets, there are several months when the ratio ($R^2_{T^+_{dp}T_o}/R^2_{T^+_{dp}E_c}$) is greater than 1, while for only one data set (October) the ratio, ($R^2_{T^+_{dp}q_i}/R^2_{T^+_{dp}E_c}$), is greater than 1. The variable T^+_{dp} seems to explain over 40% of the variance in $\dot{E}_c$. Figure 1 (c) depicts the ratio ($R^2_{q_iT_o}/R^2_{q_iE_c}$) and ($R^2_{q_iT^+_{dp}}/R^2_{q_iE_c}$) for one-month data sets and for the year long data set. There are several data sets, including the long data set, when both the ratios are greater than 1 because $R^2_{q_i,E_c}$ is generally smaller than $R^2_{T_o,E_c}$ and $R^2_{T^+_{dp},E_c}$ (less than 20%). Therefore, q_i should be dropped from the model to reduce prediction uncertainty due to collinearity. The above analysis was carried out with a building with a VAV system. The results and trends were similar.

From the above analysis it is evident that there is some collinearity between regressor variables both for the short and for the long (one year) data sets. The effect of collinearity on parameter estimates and model predictions with the long data set is probably small (with exception of q_i) for two reasons: (i) at worst, the collinearity is moderate and (ii) the characteristic of the regressor variables from year to year are almost similar. Although some short data sets had lesser collinearity between regressor variables than the long data set, the uncertainty of the model predictions can be high because the characteristics of the short data sets can be quite different as compared to a long data set.

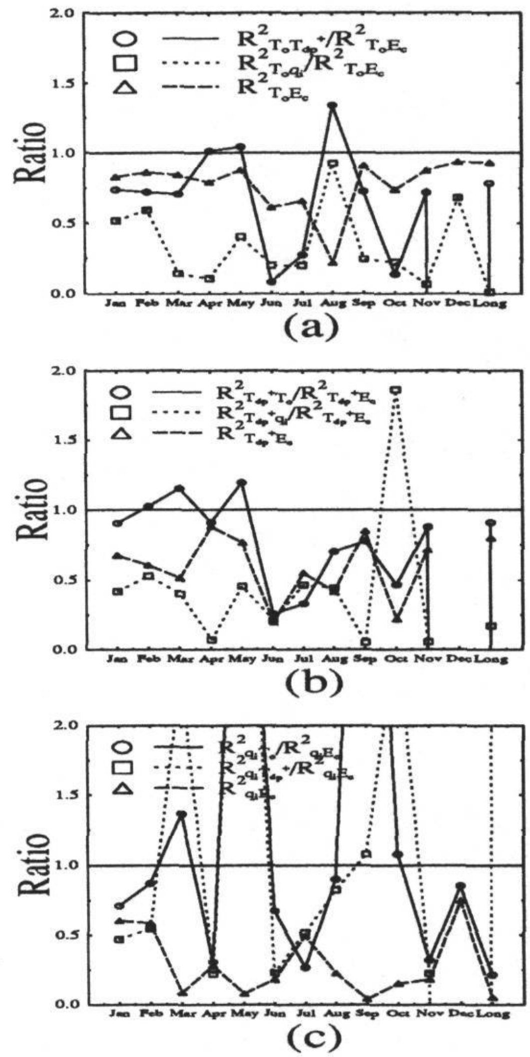

Figure 1 Ratio of the Correlation Between Independent Variables to the Correlation Between the Independent Variable and the Dependent Variable for One-Month Data Sets and a Long Data Set (EC Building With a DDCV System [09/89-08/90]). (a) T_o with T^+_{dp} and q_i, (b) T^+_{dp} With T_o and q_i, and (c) q_i with T_o and T^+_{dp}. All T^+_{dp} Values for the December data set are zero.

IDENTIFICATION OF VARIABLES THAT CONTRIBUTE TO ERRORS

In this section, we will identify the regressor variables in the MLR model that contribute to the fractional error. One of the major sources of error with the short data sets is insufficient range in the regressor variable. For example, if the functional relationship between the dependent ($\dot{E}_c$) and the independent variable (T_o) is segmented linear (4P), then there should be sufficient number of data points in both segments of the T_o range to develop a robust regression model. The regressor variable T^+_{dp} accounts for the variation in latent ventilation cooling energy use, which is small during winter months; therefore, to model this phenomena the data set should have a sufficient number of points where latent cooling occurs.

By studying the variation (range) of the regressor variables (T_o, T_{dp}, q_i, and q_{sol}) both within a short data set and over a long period, we can identify which of the variables could effect the model seasonally. For this analysis a short data set is assumed to have one month of data and a long data set to have one year's data (12 calendar months). The effect of q_{sol} on large commercial buildings with less than 20% glazed surface is small; therefore, it is dropped from further consideration. Figure 2 shows variation of $\dot{E}_c$, T_o, T_{dp}, q_i within short data sets (monthly) and from one data set to another. All quantities were measured in a large commercial building with a DDCV system. Figure 2 (a) shows, (i) minimum and maximum $\dot{E}_c$ (horizontal bars), (ii) mean $\dot{E}_c$ (square marker and a solid line), (iii) the 10^{th} and the 90^{th} percentiles for $\dot{E}_c$ within each data set, and (iv) 1^{st} and 3^{rd} quartiles for $\dot{E}_c$ within each data set.

The variation of $\dot{E}_c, T_o$, and T_{dp} within each data set is small for summer months (June, July and August), and there is more variation in the winter months and the fall months. The mean T_o values change significantly for winter to summer (variation as high as 40%). The variation in q_i (Figure 2 (d)) within each month is uniform and there no significant seasonal change.

Since the mean value of both T_o and T_{dp} shows significant seasonal variation, a short data set covering only winter or summer months may not produce a valid model. On the other hand, q_i shows no seasonal dependence, therefore, it may not effect a model based on a short data set. Figure 3 shows variation of $\dot{E}_c$ for a large commercial building with a dual-duct VAV. Unlike the range of the $\dot{E}_c$ from the DDCV system in the summer months, the range in $\dot{E}_c$ with the VAV system is wider.

DETAILED ANALYSIS OF EC SHORT DATA SETS MODELS

In this section, the reasons for differences between the model based on a short data set and the model based on a long data set are identified and analyzed. First, the EC (Engineering Center) building (1989-1990 data) with the DDCV system is analyzed. The EC building was retrofitted to a VAV system in 1991; therefore, the EC with the VAV system (1992 data) is analyzed next. The physical characteristics and operational details of the EC are given in the Appendix. The measured hourly data for the analysis included: (i) T_o, (ii) T_{dp}, (iii) q_{sol} (global horizontal solar radiation) (iv) q_i, and (v) $\dot{E}_c$, (September 1989 through August 1990). Daily values were obtained by summing the hourly values of q_{sol}, q_i and $\dot{E}_c$ over the day and the variables, T_o and T_{dp}^{+}, were averaged over the day. In general, a data set is considered short if it has fewer than three months of hourly data. For the initial analysis, the year-long data set from the EC is divided into 12 data sets, with each data set having

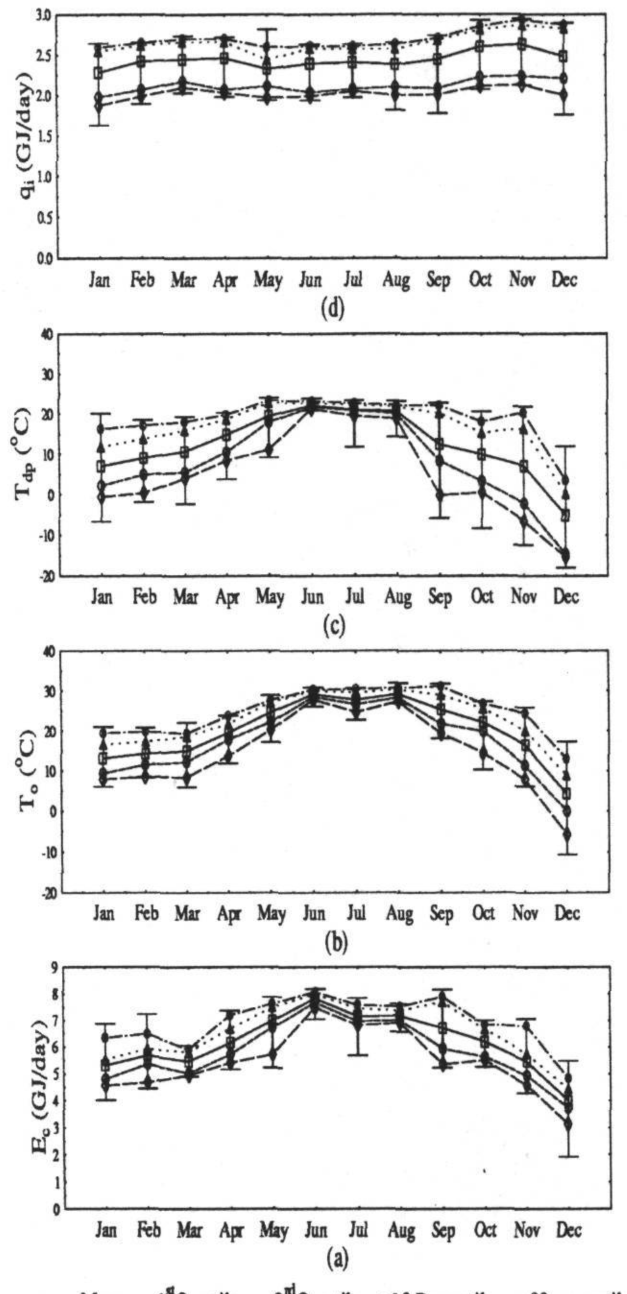

Figure 2 Variation of the mean (a) $\dot{E}_c$, (b) T_o, (c) T_{dp}^{+}, and (d) q_i for One-Month Data Sets From the EC Building With a DDCV System (9/89-8/90).

one calendar month of daily data.

One-Month Data Sets: DDCV System

The cooling energy consumption in the EC building did not show a segmented linear relationship with T_o, because the change

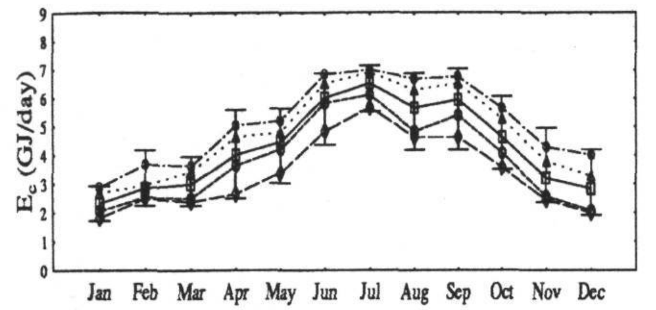

Figure 3 Variation of the mean $\dot{E}_c$ for One-Month Data Sets From the EC Building With a VAV System (1992).

in the hot deck temperature with T_o was small. Therefore, the terms I and $I * T_o$ were dropped while developing the regression models. First, twelve daily regression models were developed for each of the 12 one-month data sets and then each of the 12 models were used to predict the average daily $\dot{E}_c$ for the entire year. Figure 4 shows $\dot{E}_c$ as a function of T_o for the 12 one-month data sets along with the residuals (measured - predicted) for the entire year determined using the model based on that months data set.

All one-month data sets, with the exception of the October data set, show biased residuals. The residuals vary from a constant negative for the January and the February data sets, increasing with T_o for March, August and December data sets, and decreasing with T_o for June, July and September data sets.

To study qualitatively the difference in the prediction of $\dot{E}_c$ from models based on a short data set and a model based on a long data set as a function of T_o, the effects of T_{dp}^+ and q_i have to be isolated. The effect of T_o can be isolated by assuming T_{dp}^+ as zero and q_i is assumed to constant. With these modifications the model is given by:

$$\dot{E}_{c,s} = \alpha + \beta_0 T_o + \beta_4 \overline{q_i} \tag{4}$$

where $\overline{q_i}$ is the average internal gains value. Similarly, the effect of T_{dp}^+ can be qualitatively analyzed by using the following model:

$$E_{c,s} = \alpha + \beta_0 \overline{T_o} + \beta_3 T_{dp}^+ + \beta_5 \overline{q_i} \tag{5}$$

Figure 5 enables comparison of selected models from one-month data sets with the model from the long data set using Eqs. 4 and 5. Figure 5 (a), (b), and (c) shows the effect of T_o and Figure 5 (d), (e), and (f) shows the effect of T_{dp}^+. The residuals in Figure 4 for the January data set are mostly negative and show no relationship with T_o. The model of $\dot{E}_c$ as a function of T_o for January (Figure 5(a)) compares well with the model based on the long data set, but the model of $\dot{E}_c$ as a function of T_{dp}^+ predicts slightly larger values than the long data set model (Figure 5(d)). Therefore, the residuals are negative with the January model. For the March data set the residuals increase with T_o, because the March data set has a smaller slope than the long data set (Figure 5 (a)) and the slope of T_{dp}^+ from the March data set is almost zero.

The residuals from the June data set are decreasing with T_o, while they are increasing for the August data set. The contribution of T_{dp}^+ from the June data set differs only slightly from the long data set, while the model using T_o from the June data set is significantly different from the long data set which has a larger slope; therefore, the residuals are decreasing with T_o. The slope of T_o for the August data set is small. Therefore, the residuals are increasing with T_o.

The residuals for the October data set are unbiased (Figure 4), because the relationship of $\dot{E}_c$ with both T_o and T_{dp}^+ is similar to those of the long data set (Figure 5 (c) and (f)). The relationship of $\dot{E}_c$ with T_o for the December data set is similar to that with the long data set, but the contribution of T_{dp}^+ is zero because the outdoor dew-point temperature in December was always lower than the average surface temperature of the cooling coil (i.e., no latent cooling).

One-Month Data Sets: VAV System

In this section, we present a detailed analysis of the difference between the cooling energy consumption model for a VAV system based on a short data set and the model based on a long data set. Unlike the cooling energy consumption with the DDCV system, the cooling energy consumption with the VAV system in the EC building shows a segmented linear relationship with T_o. Therefore, using Eq. 1 the daily regression models were developed for the 12 one-month data sets and then each of the 12 models were used to predict the daily $\dot{E}_c$ for the entire year. Figure 6 shows $\dot{E}_c$ as a function of T_o for the 12 one-month data sets along with the residuals (measured - predicted) for the entire year.

All models based on one-month data sets show bias in the residuals. Unlike the DDCV residuals, the residuals from the VAV system show a segmented linear relationship with T_o. Since the air flow rate is modulated in a VAV system to match the zone thermal load, the cooling energy is a stronger function of the outdoor conditions (T_o and T_{dp}^+). The VAV system in the EC building reaches the minimum flow condition around 21 ^oC outdoor dry-bulb temperature, i.e. below 21 ^oC the air flow is no longer modulated. Therefore, at this outdoor temperature there is a change in the slope of T_o. If the data sets contain data either above or below 21 ^oC only, then the model is totally inadequate to predict consumption on an annual basis.

To study qualitatively the difference in the prediction of $\dot{E}_c$

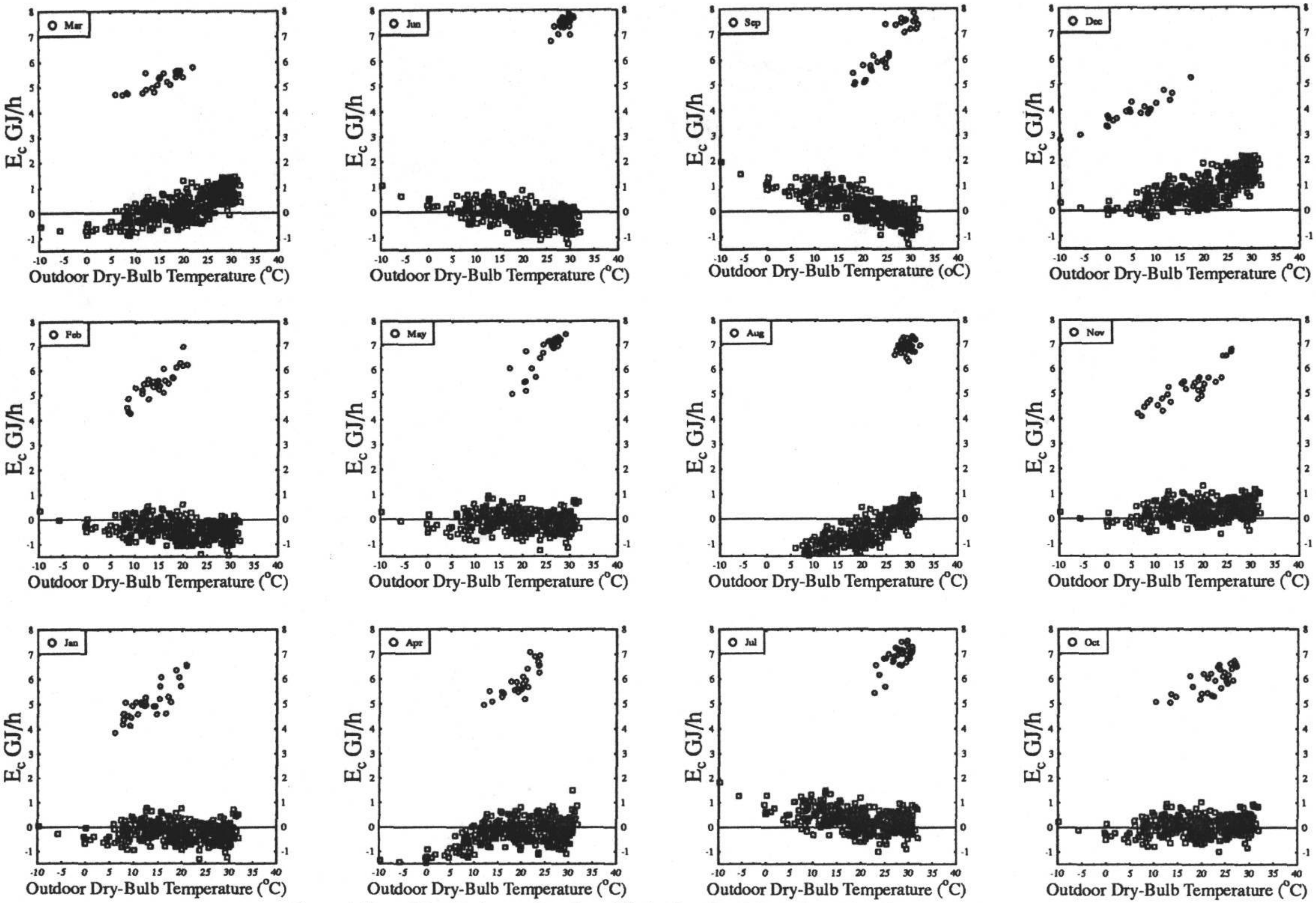

Figure 4 E_c and Residuals as a Function of T_o for One-Month Data Sets; "Jan" Represents the January Data Set, "Feb" the February Data Set, and so on. (EC Building With the DDCV System (1989-1990))

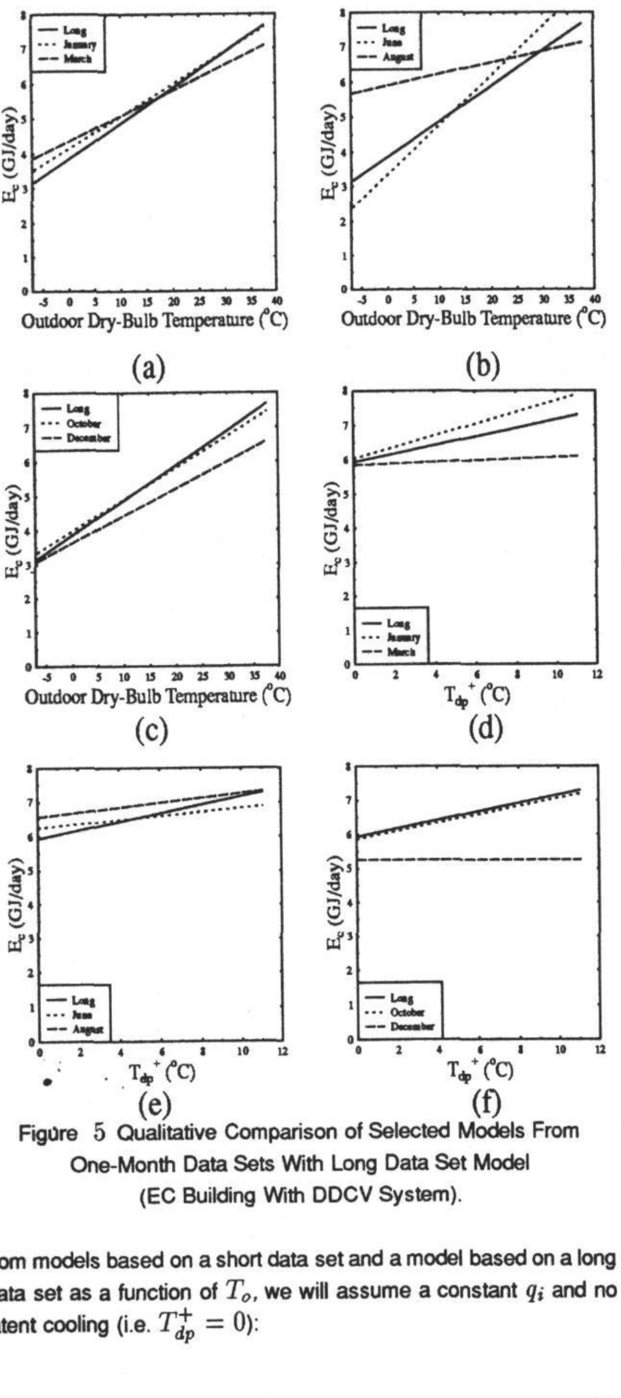

Figure 5 Qualitative Comparison of Selected Models From One-Month Data Sets With Long Data Set Model (EC Building With DDCV System).

from models based on a short data set and a model based on a long data set as a function of T_o, we will assume a constant q_i and no latent cooling (i.e. $T_{dp}^+ = 0$):

$$\dot{E}_c = \alpha + \beta_0 T_o + \beta_1 I + \beta_2 I T_o + \beta_5 \overline{q_i} \qquad (6)$$

Similarly, the effect of T_{dp}^+ can be qualitatively analyzed by using the following model:

$$E_{c,s} = \alpha + \beta_0 \overline{T_o} + \beta_3 T_{dp}^+ + \beta_5 \overline{q_i} \qquad (7)$$

Figure 7 permits a comparison of selected models from one-month data sets with the model from the long data set using Eqs. 6 and 7. Figure 7 (a), (b), and (c) shows the effect of T_o and Figure 7 (d), (e), and (f) shows the effect of T_{dp}^+. The residuals in Figure 6 for the January data set show a 3P-like pattern. For low outdoor dry-bulb temperatures (below 10 ^oC), $\dot{E}_c$ prediction from the January data set is identical to that from the long data set (Figure 7 (a)). However, for T_o greater than 18 ^oC the residuals for the January data set are increasing with T_o (Figure 6); this effect is clearly evident from Figure 7 (a). The VAV system in the EC reaches its minimum flow condition at 21 ^oC; therefore, there is a change in the slope above 21 ^oC outdoor temperature. Since the temperatures in the January data set are between 0 and 15.6 ^oC, the January regression model is not able to capture the change in the slope at around 21 ^oC.

Since the outdoor dry-bulb temperatures in the April data set are between 10 ^oC and 21 ^oC, the model is doing poorly at both extremes. In addition, the April model is predicting a higher contribution from T_{dp}^+ than the long data set model. These combined effects are the cause for the biased residuals (April Figure 6). The slope of T_{dp}^+ with the June data set is nearly identical to that from the long data set model. Above 21 ^oC T_o, the contribution from the June model are identical to the long data set, but the model under predicts increasingly for T_o below 21 ^oC.

The range in T_o for the August data set is small; so the change-point behavior is not captured. The contributions from T_{dp}^+ from the October and the December data set models are similar to that from the long data set model. Since, the temperature range in these two data sets is mostly below 21 ^oC, these models do not show the higher slope above 21 ^oC and are under predicting above T_o of 21 ^oC.

The regression models based on one month data, for both the DDCV and the VAV systems, are inadequate in predicting the annual energy consumption. For buildings with minimum outdoor air intake (10-20%), the cooling energy consumption with a DDCV system is not a strong function of the outdoor conditions, because of constant air flow rate. However, for buildings with 100% outdoor air intake (medical facilities and laboratories buildings) it is strong. Since the EC building only takes in 15% outdoor air, the residuals from the one-month data sets are moderate. Also the use of the hot deck reset schedule will induce segmented linear behavior in the cooling energy consumption as a function of T_o.

In contrast, the cooling energy consumption with a VAV system is a strong function of the outdoor conditions, because the air flow rate is modulated to meet the thermal load of the building. The air flow is modulated until it reaches a certain minimum (30-50% of the rated capacity), below which the air flow rate is a

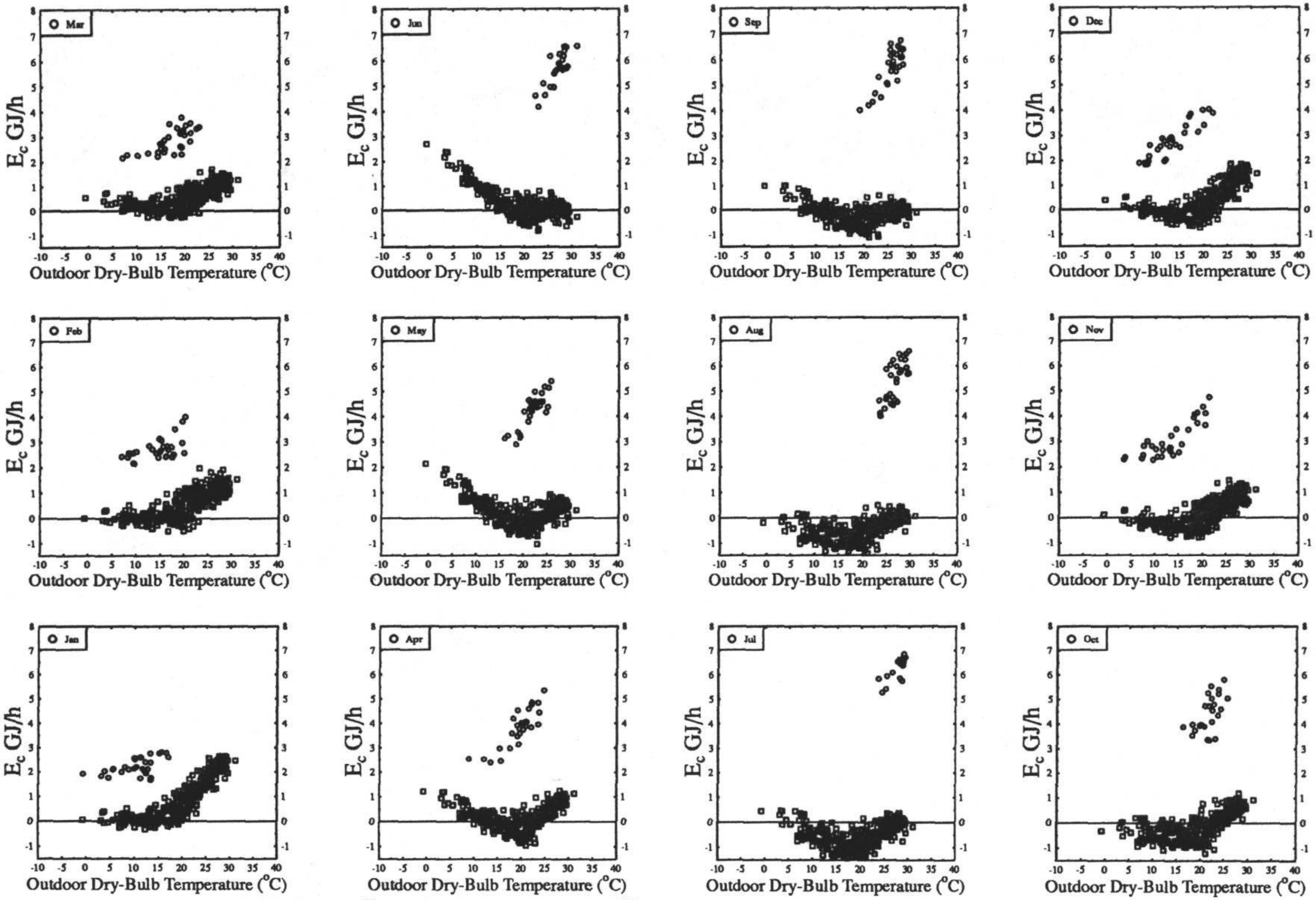

Figure 6 E_c and Residuals as a Function of T_o for One-Month Data Sets; "Jan" Represents the January Data Set, "Feb" the February Data Set, and so on. (EC Building With the VAV System (1992))

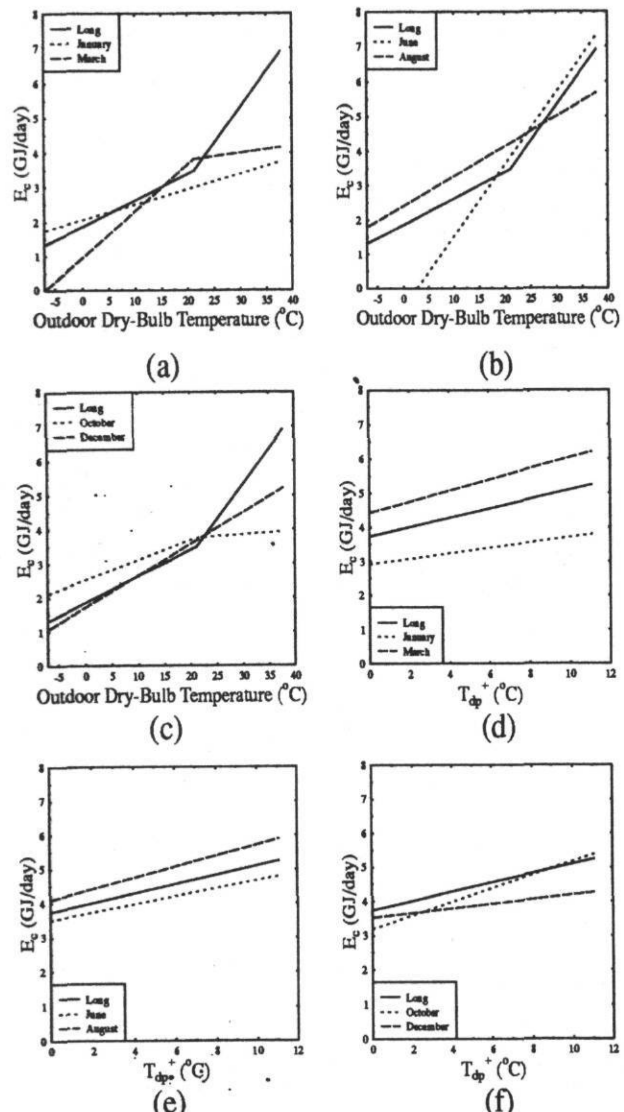

Figure 7 Qualitative Comparison of Selected Models From One-Month Data Sets With Long Data Set Model (EC Building With VAV System).

constant. Therefore, there is a change in the slope of T_o when the minimum flow condition is reached. This makes the cooling energy consumption segmented linear as a function of T_o. Therefore, the one-month data sets of the EC building with the VAV system have highly biased residuals.

QUANTIFYING THE ERROR FROM SHORT DATA SETS

Five buildings for which a full year of data were available were chosen for the analysis. The physical characteristics of the selected buildings are shown in the Appendix. These buildings are large commercial/institutional buildings located in Central Texas. Two buildings (EC, and WEL) have DDCV systems and three buildings (EC, BUR, and WIN) have VAV systems. A major fraction of the conditioned area in WEL is made up of laboratories; therefore, it takes in over 80% of outdoor air.

One way to quantify the annual energy prediction error is to calculate the MBE (Mean Bias Error):

$$MBE = \frac{\sum_{j=1}^{j=n}(E_{c,j} - \hat{E}_{c,j})}{n\overline{E}_c} \tag{8}$$

where n represents the total number of days in the data set. $E_{c,j}$ is the measured daily cooling energy use, $\hat{E}_{c,j}$ is the predicted cooling energy use and $\overline{E_c}$ is the yearly average daily cooling energy use.

Figure 8 shows the variation of MBE as a function of average monthly outdoor dry-bulb temperature for two buildings with DDCV systems. The MBE in the WEL building is a much stronger function of T_o, varying from +30% at 4 ^oC to -15% at 32 ^oC. The MBE for the EC building, however, only varies from $\pm$8%. Unlike the EC building the WEL building takes in about 80% fresh air and the hot deck temperature is also reset based on T_o; therefore, the MBE shows such a strong relationship with T_o.

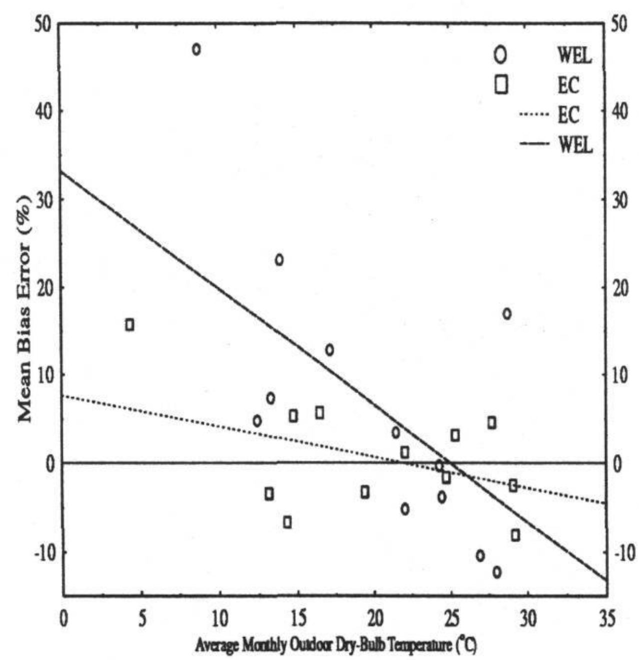

Figure 8 MBE as a Function of Average Monthly Temperature for the One-Month Data Sets (DDCV). Individual Data Points as Well as Corresponding Regression Lines Are Shown.

Figure 9 shows variation of MBE as a function of average monthly outdoor dry-bulb temperature for three buildings with VAV systems. On an average the MBE varies from +35% at 4 ^oC to -10% at 32^oC. Since all three buildings are operated identically, the MBE appears to similar.

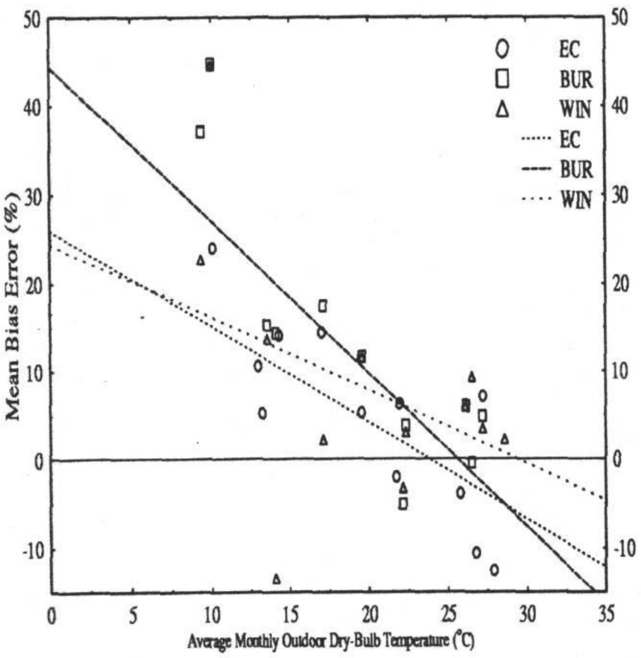

Figure 9 MBE as a Function of Average Monthly Temperature for the One-Month Data Sets (VAV). Individual Data Points as Well as Corresponding Regression Lines Are Shown.

CONCLUSIONS

Certain issues relating to regression models based on short data sets have been investigated in this study. Reasons for differences between the models based on short data sets and those based on long data sets have also been presented.

Since mean monthly values of T_o and T_{dp} vary significantly (50%) over a season, a short data set covering only the winter or the summer months may produce an inadequate model. The MBE for a building with a DDCV system which takes in 80% outdoor air varied from +30% at 4 ^oC to -15% at 32 ^oC. On the other hand the MBE for a building with a DDCV system which takes in only 15% outdoor air had a MBE variation of only $\pm$8%. The MBE for all three buildings with VAV systems varied from +35% at 4 ^oC to -10% at 32 ^oC.

Since the air flow is modulated in a VAV system, the $\dot{E}_c$ is a segmented-linear function of T_o; therefore, if a data set does not have sufficient number of points covering the entire range the uncertainty from model predictions will be high.

ACKNOWLEDGMENTS

This work was supported by the Texas State Energy Conservation Office as part of the LoanSTAR monitoring and analysis program.

REFERENCES

Draper, N., Smith, H., 1981. Applied Regression Analysis, 2nd Edition, John Wiley & Sons, New York, NY.

Fels, M. (Ed), 1986. "Special Issue Devoted to Measuring Energy Savings, The Princeton Scorekeeping Method (PRISM)," Energy and Buildings, vol. 9, nos. 1 and 2.

Kaplan, M., McFerran, J., Jansen, J., and Pratt, R. 1990. "Reconciliation of a DOE2.1c Model with Monitored End-Use Data for a Small Office Buildings," ASHRAE Transactions, vol. 96, pt. 1.

Katipamula, S., and Claridge, D.E., 1992. "Monitored Air Handler Performance and Comparison With A Simplified System Model," ASHRAE Transactions, vol. 98, pt. 2, pp. 341--351.

Katipamula, S., and Claridge, D.E., 1993. "Use of Simplified System Models to Measure Retrofit Energy Savings," ASME Journal of Solar Energy Engineering, vol. 116, no. 2, May, pp. 57-68.

Katipamula, S., Reddy, T.A., and Claridge, D.E., 1994a. "Development and Application of Regression Models To Predict Cooling Energy Consumption in Large Commercial Buildings," Proceedings of the ASME/JSME/JSES International Solar Energy Conference, San Francisco, California, March 27 -- 30, pp. 299--306.

Katipamula, S., Reddy, T.A., and Claridge, D.E., 1994b. "Effect of Time Resolution on Statistical Modeling of Cooling Energy Use in Large Commercial Building," Submitted for publication in the ASHRAE Transactions.

Kissock, J.K., J.S. Haberl, Claridge, D.E., and Reddy, T.A., 1992. "Measuring Retrofit Savings for the Texas LoanSTAR Program: Preliminary Methodology and Results," Proceedings of the ASME/JSES/KSES International Solar Energy Conference, Maui, Hawaii, pp. 299--308.

Kissock, J.K., Reddy, T.A., Fletcher, D., and Claridge, D.E., 1993. "The Effect of Short Data Periods on The Annual Prediction Accuracy of Temperatures-Dependent Regression Models of Commercial Buildings Energy Use," Proceedings of the ASME/ASES/SED International Solar Energy Conference, Washington, D.C., April.

MacDonald, J.M., and Wasserman, D.M., 1988, "Investigation of Metered Data Analysis Methods for Commercial and Related Buildings," ORNL/CON-279, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Mullet, G.M., 1976. "Why Regression Coefficients Have the Wrong Sign," Journal of Quality Technology, Vol. 8, no. 3, July, pp. 121 -- 126.

Rachlin, J., Fels, M., and Socolow, R., 1986. "The Stability of PRISM Estimates," Energy and Buildings, vol. 9, nos. 1 and 2.

Reddy, T.A., Kissock, K.S., Katipamula, S., and Claridge, D.E., 1994a. " An Energy Efficiency Index to Evaluate Simultaneous Heating and Cooling Effects in Large Commercial Buildings," ASME Journal of Solar Energy Engineering Vol. 116, no. 2, May, pp. 79-87.

Reddy, T.A., Katipamula, S., and Claridge, D.E., 1994b. "Functional Basis Of Thermal Energy Use," Accepted for publication in the ASME Journal of Solar Energy Engineering.

Ruch, D., and Claridge, D.E., 1992. "A Four Parameter Change Point Model for Predicting Energy Consumption in Commercial Buildings," ASME Journal of Solar Energy Engineering Vol. 114, pp. 77.

Weisberg, S., 1976. Applied Linear Regression, Second Edition, John Wiley and Sons, New York, NY.

NOMENCLATURE

$= \alpha, \beta_0, \beta_1, \beta_2, \beta_3,$

$=$ β_4, β_5 are regression coefficients

$=$ all terms with a bar over them indicate average values

$=$ all terms with a "prime" indicate parameters from short data set

$\dot{E}$ $=$ rate of energy use (Btu/unit time)

I $=$ indicator variable

T $=$ dry-bulb temperature (F)

T_{dp} $=$ outdoor dew-point temperature (F)

T_o $=$ outdoor dry-bulb temperature (F)

$\overline{T_r}$ $=$ average outdoor dry-bulb temperature to right of the change point (F)

T_s $=$ cooling coil leaving air dry-bulb temperature (F)

q_i $=$ internal gains from lights and equipment (Btu/unit time)

q_{sol} $=$ horizontal global solar radiation (Btu/h-sf)

APPENDIX

Table A1 Building Characteristics.

Building /Type	Floor Area (m ^{2})	Area	Construction	Glazing % of Area	Glazing Type	Outdoor o/a Intake (%)	Supply Temp. (^oC) T_c / Max. T_h	Building Operation People/HVAC (hrs/day)
EC/1 (DDCV)	30,150	3,810	Insulated Cement Block	22	Sinlge Pane	10-20	15.6/40.6	12/24
EC/1 (VAV)	30,150	3,810	Insulated Cement Block	22	Sinlge Pane	10-20	12.8/40.6	12/24
WEL/1	40,850	6,230	Red Face Brick on Block	17	Single Pane	90	12.8/35	12/24
BUR/2	9,610	3,250	Pre-Cast Cement Block	7	Single Pane	10-15	12.8/35	12/24
WIN/2	10,130	4,000	Cement Block with Face Brick	10	Single Pane	10-15	12.8/35	12/24

1:Class/Lab/Office; 2:Class/Office