Monash Electricity Forecasting Model

Professor Rob J Hyndman

B.Sc. (Hons), Ph.D., A.Stat.

Dr Shu Fan

B.S., M.S., Ph.D.

Business & Economic Forecasting Unit

Telephone: (03) 9905 2358
Fax: (03) 9905 5474

ABN: 12 377 614 012

Version 2015.1

28 May 2015

Contents

1 Introduction

Monash University has been working with various Australian electric power system industry bodies to develop better forecasting techniques for electricity demand since 2007. The Monash Electricity Forecasting Model (MEFM) was developed to forecast the probability distribution of electricity demand, and the underlying theory and methodology has been published in a leading academic journal (Fan and Hyndman, 2010; Hyndman and Fan, 2010). The model has been used to forecast the probability distribution of annual, seasonal and weekly peak electricity demand and energy consumption for various regions of Australia, including the regions of the National Electricity Market (NEM), the SWIS of Western Australia and the Keilor Terminal Station in Victoria. Short-term forecasting software has also been developed to forecast half-hourly electricity demand (up to seven days ahead) in South Australia and Victoria.

The MEFM, as shown in Figures 1 and 2 is unique in several respects. First it allows forecasts of short-, medium- and long-term demand by carefully modeling the dynamic non-linear relationships between the various driver variables and electricity usage, and then allowing for long-term changes in the driver variables as a result of economic, demographic and technological development, as well

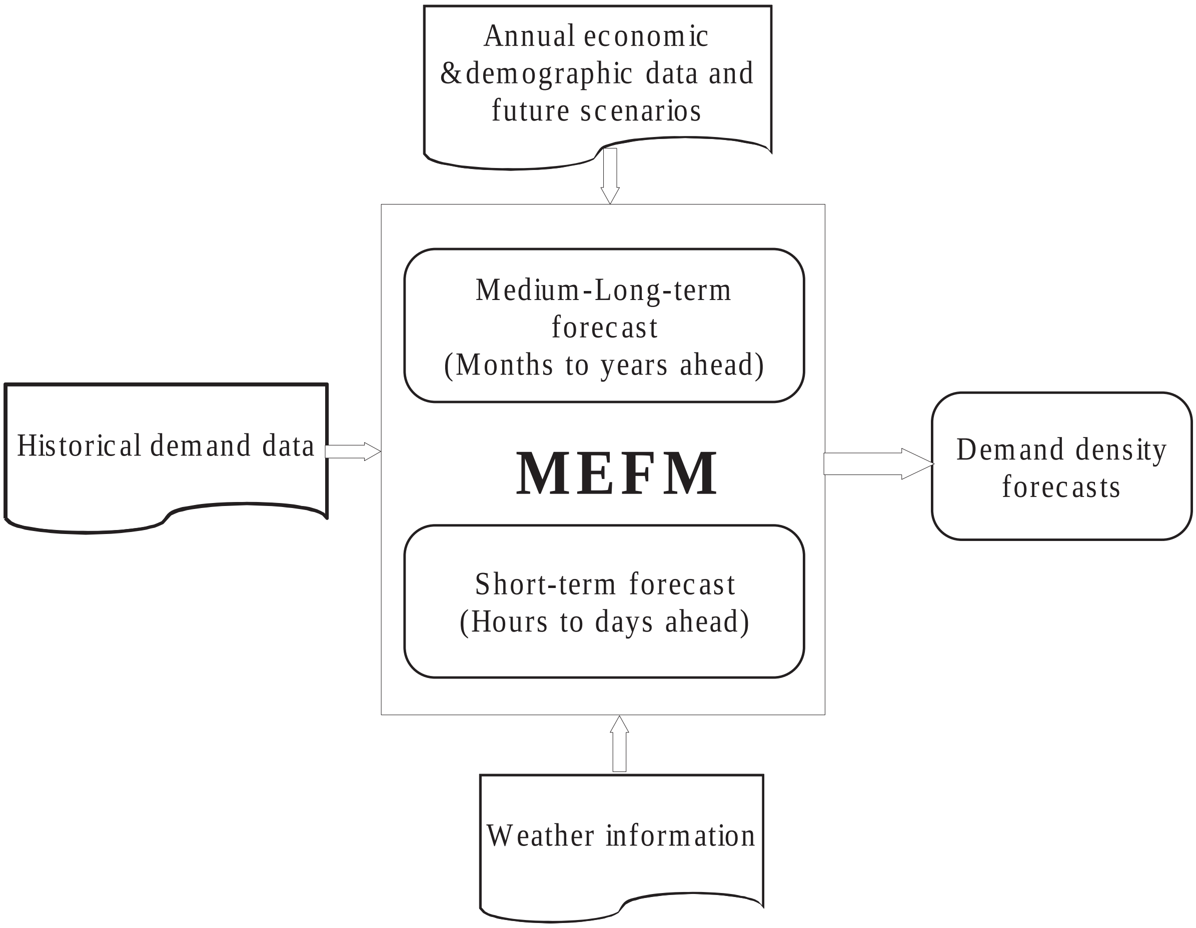
Annual economic & demographic data and future scenarios
Medium-Long-term forecast (Months to years ahead)
Historical demand data
MEFM
Demand density forecasts
Short-term forecast (Hours to days ahead)
Weather information

Figure 1: *Monash Electricity Forecasting Model.*

as climate change. No other approach has attempted to combine all of these relationships into one coherent model. Second, the model is designed to capture the entire probability distribution of demand, whereas almost all previous work has focused on average demand. As a result, we can obtain forecasts of peak demand which are important when planning sufficient capacity, and accurate short-term forecasts which are essential for system scheduling and the electricity market.

The forecasting methodological framework is illustrated in Figure 2, and can be summarized in three stages: (1) modeling; (2) simulating and forecasting; and (3) evaluating.

First of all, the relationships between demand and the driver variables, including temperatures, calendar effects and some demographic and economic variables, are estimated using semi-parametric additive models (Ruppert, Wand, and Carroll, 2003). The model can be split into long-term annual or seasonal effects (driven by economic and demographic variables) and half-hourly effects (driven by temperature and calendar variables), with the two parts of the model estimated separately. The input variables of the annual/seasonal model are selected using Akaike's Information Criterion (AIC) and

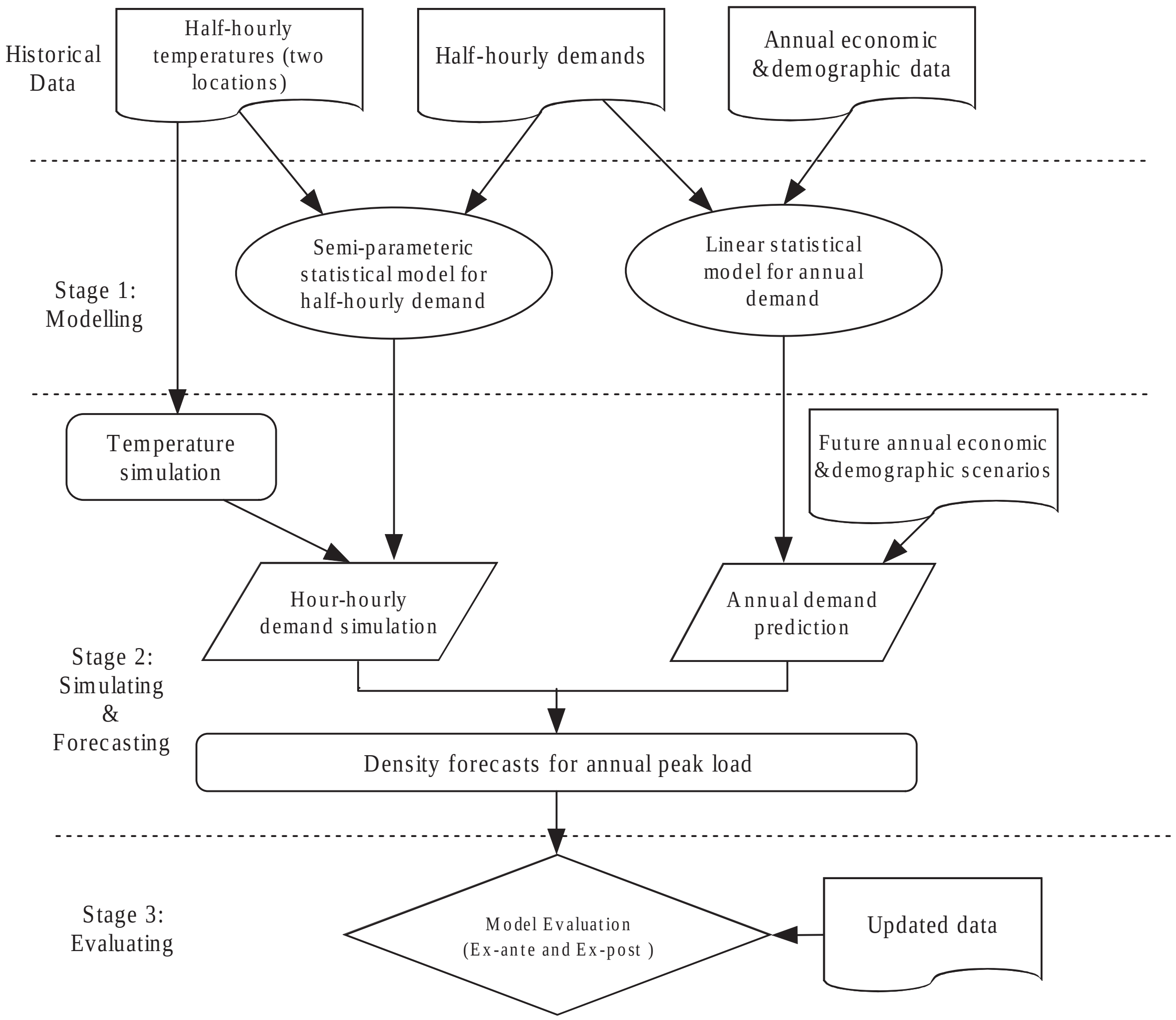
Historical Data
Half-hourly temperatures (two locations)
Half-hourly demands
Annual economic &demographic data
Stage 1: Modelling
Semi-parameteric statistical model for half-hourly demand
Linear statistical model for annual demand
Temperature simulation
Future annual economic &demographic scenarios
Hour-hourly demand simulation
Annual demand prediction
Stage 2: Simulating & Forecasting
Density forecasts for annual peak load
Stage 3: Evaluating
Model Evaluation (Ex-ante and Ex-post)
Updated data

Figure 2: *Block diagram of the Monash Electricity Forecasting Model.*

the input variables of the half-hourly model are selected by minimizing the out-of-sample forecasting errors on a cross-validation sample.

In the second stage, the forecast distributions are obtained from the estimated model using a mixture of temperature and residual simulations, and future assumed demographic and economic scenarios. A seasonal bootstrapping method with variable blocks is applied to resample the residuals and temperatures. The temperature bootstrap is designed to capture the serial correlation that is present in the data due to weather systems moving across different regions. The residual bootstrap captures the remaining serial correlation that has not been modelled with the predictor variables.

Finally, to evaluate the forecasting performance of the model, the actual demand of a summer with two different types of predictions (ex ante forecasts and ex post forecasts) are compared. Specifically, **ex ante forecasts** are the forecasts made in advance using whatever information is available at the time. On the other hand, **ex post forecasts** are those that are made using information on the "driver variables" that is only known after the event being forecast. The difference between the ex ante forecasts and ex post forecasts provides a measure of the effectiveness of the model for forecasting (taking out the effect of the forecast errors in the input variables).

2 Semi-parametric demand model

We use semi-parametric additive models to estimate the relationship between demand and the driver variables. The model is in the regression framework but with correlated errors. This is similar to the models developed by others including McSharry, Bouwman, and Bloemhof (2005) and Ramanathan et al., 1997, but we propose a large number of modifications and extensions to make the models more general, robust and effective. In particular, our models allow nonlinear and nonparametric terms using the framework of additive models (Hastie and Tibshirani, 1995).

We fit a separate model to the data from each half-hourly period. The demand patterns change throughout the day and better estimates are obtained if each half-hourly period is treated separately. Ramanathan et al. (1997) also proposed a separate regression model for each time of the day (they used hourly data).

In addition, we allow temperature and day-of-week interactions by modelling the demand for workdays and non-workdays (including weekends and holidays) separately. We also evenly split the data into 3 subsets: morning, afternoon and evening and select the best model for each subset separately, by doing this, we reduce the computational burden while having the model variables optimized for different periods during the day.

Specific features of the models we consider are summarized below:

- temperature effects are modelled using regression splines;
- temperatures from the last three hours and the same period from the last six days are included;
- economic and demographic variables are modelled linearly (on a log scale);
- errors are correlated.

We use three different indexes for time:

1. p denotes the time of day, from $p = 1$ (meaning 12 midnight – 12:30am) to $p = 48$ (meaning 11:30pm – 12 midnight);
2. t denotes the time period in half-hourly intervals, from $t = 1$ meaning the first half-hourly period for which we have observed demand data, and increasing throughout the entire data set. After 15 years of half-hourly data, we will have approximately $t = 48 \times 365.25 \times 15 = 262980$, depending on where leap years fall;
3. i denotes the time period in seasonal intervals (usually quarterly), from $i = 1$ meaning the first seasonal period for which we have observed demand data, and increasing throughout the data set. For example, after 15 years of data, $i = 4 \times 15 = 60$.

Thus $p = [(t-1) \mod 48] + 1$ and, for quarterly data, $i \approx \lfloor t/(365.25/4 \times 48) \rfloor$. Here, $\lfloor u \rfloor$ is the largest integer less than u. (The approximation arises due to unequal length quarters.) We switch between these indexes depending on which is most convenient, and sometimes we use two of the indexes simultaneously.

We assume that any major industrial loads (such as required by large mines or smelters) are not affected by ambient temperatures, and we remove this portion of demand from the data — it is be modelled and forecasted separately from the methods described here. Let $y_{t,p}$ denote the *remaining* demand at time t (measured in half-hourly intervals), coinciding with period p $(p = 1, \dots, 48)$.

One feature is that the model has been split into two separate models, one model based on the seasonal demographic, economic variables, and degree days (all linear terms), and the other based on the remaining variables which are measured at half-hourly intervals. Thus,

$$\log(y_{t,p}) = \log(y^*_{t,p}) + \log(\bar{y}_i)$$

where $\bar{y}_i$ is the average demand for season i (e.g., quarter) in which time period t falls, and $y^*_{t,p}$ is the normalized demand for time t and period p. For example, Figure 3 shows data for the summer season of Victoria, with the top panel showing the original demand data and the average seasonal demand values shown in red, and the bottom panel showing the half-hourly normalized demand data. Then

$$\log(y^*_{t,p}) = h_p(t) + f_p(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t}) + e_t \tag{1}$$

$$\text{and} \qquad \bar{y}^{pc}_i = \sum_{j=1}^{J} c_j z_{j,i} + \varepsilon_i, \tag{2}$$

where

- $\bar{y}^{pc}_i = \bar{y}_i / P_i$ is the per-capita seasonal average demand;
- P_i is the population in season i;
- $h_p(t)$ models all calendar effects;
- $f_p(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t})$ models all temperature effects where $\boldsymbol{w}_{1,t}$ and $\boldsymbol{w}_{2,t}$ are a vectors of recent temperatures at two locations;
- $z_{j,i}$ is a variable in season i (e.g., economic variables); its impact on demand is measured via the coefficient c_j (note that these terms do not depend on the period p);
- n_t denotes the model error at time t.

Per-capita demand (demand/population) and per-capita GSP (GSP/population) are considered in the seasonal model, so as to allow for both population and economic changes in the model. For half-hourly demand, the log value is modelled, rather than the raw demand figures. We tried a variety of transformations of demand from the Box and Cox (1964) class and found that the logarithm resulted in the best fit to the available data. Natural logarithms have been used in all calculations.

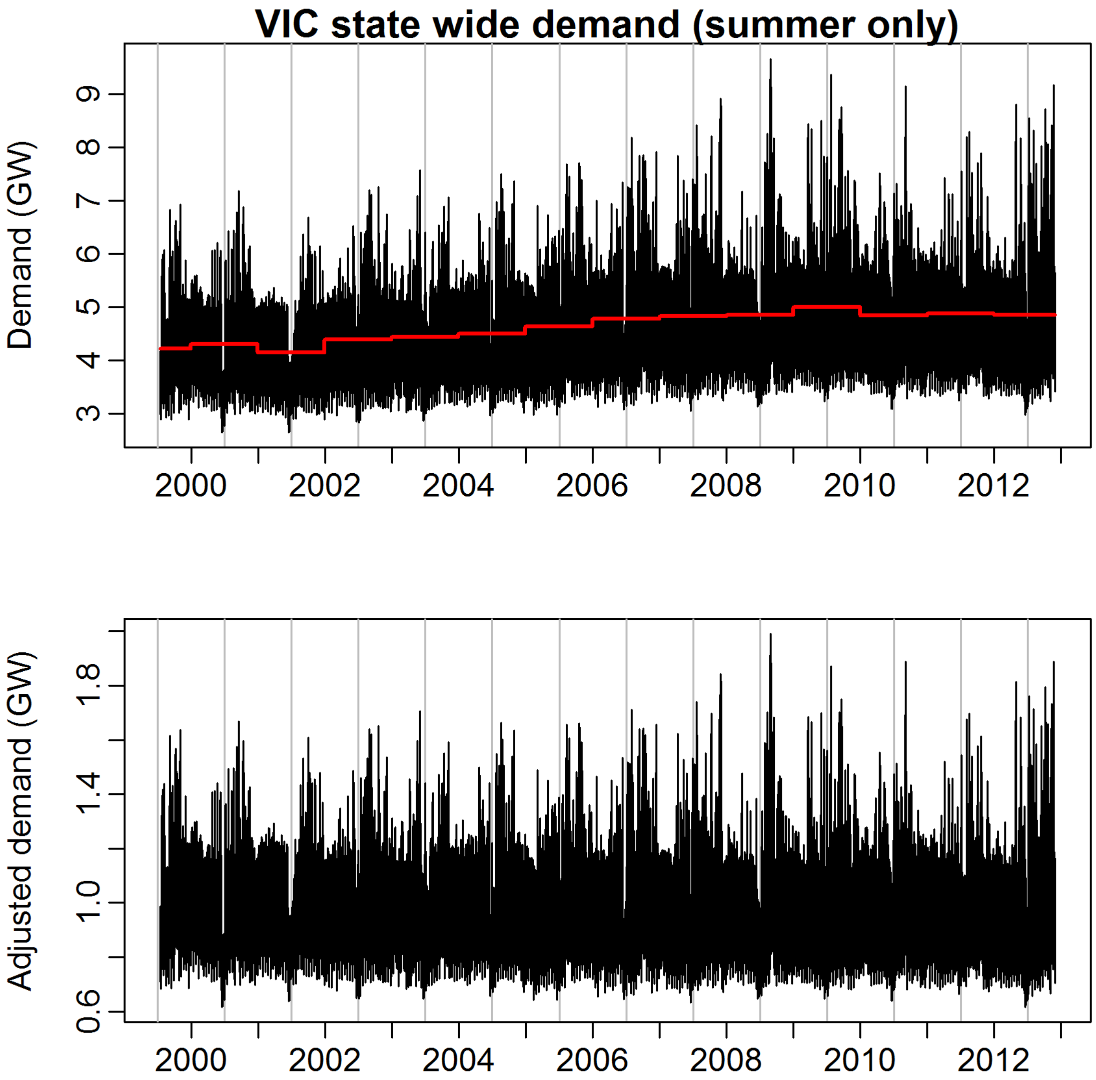

Figure 3: *Top: Half-hourly demand data for Victoria from 2000 to 2013. Bottom: Adjusted half-hourly demand where each year of demand is normalized by seasonal average demand. Only data from October–March are shown.*

Combining (1) and (2), we can see that the model for each half-hour period can be written as

$$\log(y_{t,p}) = h_p(t) + f_p(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t}) + \log(P_i) + \log\left(\sum_{j=1}^{J} c_j z_{j,i}\right) + n_t. \tag{3}$$

The error term n_t will be serially correlated, reflecting the fact that there are other environmental conditions that are not captured in this model.

2.1 Calendar effects

$h_p(t)$ includes seasonal, weekly and daily seasonal patterns as well as public holidays:

$$h_p(t) = \alpha_{t,p} + \beta_{t,p} + \gamma_{t,p} + \ell_p(t) \tag{4}$$

- $\alpha_{t,p}$ takes a different value for each day of the week (the "day of the week" effect);
- $\beta_{t,p}$ takes value zero on a non-work day, some non-zero value on the day before a non-work day and a different value on the day after a non-work day (the "holiday" effect);
- $\gamma_{t,p}$ takes value zero except for the time period from 12 midnight to 1am on New Year's Day (the "New Year's Eve" effect);
- $\ell_p(t)$ is a smooth function that repeats each year (the "time of summer" effect).

The smooth function $\ell(t)$ is estimated using a cubic regression spline. Knots are chosen automatically as described later in this report.

2.2 Temperature effects

The function $f_p(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t})$ models the effects of recent temperatures on the normalized demand. Because the temperatures at the two locations are probably highly correlated, we do not use these directly. Instead, we take the average temperature across the two sites

$$x_t = (w_{1,t} + w_{2,t})/2$$

and the difference in temperatures between the two sites

$$d_t = (w_{2,t} - w_{1,t}).$$

These will be almost uncorrelated with each other making it easier to use in statistical modelling. Then the temperature effects are included using the following terms:

$$\begin{aligned} f_p(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t}) = \sum_{k=0}^{6} \big[f_{k,p}(x_{t-k}) + g_{k,p}(d_{t-k}) \big] + \sum_{j=1}^{6} \big[F_{j,p}(x_{t-48j}) + G_{j,p}(d_{t-48j}) \big] \\ + q_p(x_t^+) + r_p(x_t^-) + s_p(\bar{x}_t), \end{aligned} \tag{5}$$

where

- x_t^+ is the maximum of the x_t values in the past 24 hours;

- x_t^- is the minimum of the x_t values in the past 24 hours;
- $\bar{x}_t$ is the average temperature in the past seven days.

Each of the functions ($f_{k,p}$, $g_{j,p}$, $F_{k,p}$, $G_{j,p}$, q_p, r_p and s_p) is assumed to be smooth and is estimated using a cubic regression spline.

2.3 Demographic and economic effects

The remaining terms in the model are the seasonal variables which enter via the summation term

$$\sum_{j=1}^{J} c_j z_{j,i}. \tag{6}$$

Typically, the variables $z_{1,i}, \dots, z_{J,i}$ include a measure of the strength of the local economy (e.g., Gross State Product), a measure of the price of electricity, the number of cooling degree days and heating degree days for season i, and any other variable measured on a seasonal scale that might be relevant.

Cooling degree days and heating degree days are defined respectively as

$$\begin{aligned} z_{\text{CDD},i} &= \sum_j \max(0, \bar{T}_{j,i} - 18.5) \\ z_{\text{HDD},i} &= \sum_j \max(0, 18.5 - \bar{T}_{j,i}) \end{aligned}$$

where $\bar{T}_{j,i}$ is the daily mean temperature for day j in season i and the sums are over all days of the season. These provide a measure of how hot (cold) a given season is overall. The thresholds (given as 18.5 in each equation above) are roughly the temperatures at which a cooler or heater is turned on. It is not necessary to use the same threshold in each equation, and the above values vary for different regions.

Because economic relationships are usually relatively weak, and change slowly, we chose to estimate each of these as a linear term with the same coefficient for each time period. Furthermore, the economic variables are subject to substantial measurement error. Consequently, more complicated relationships for these variables are not justified.

2.4 Regression splines

We use regression splines to model the non-linear temperature relationships and also the relationship between demand and the time of year. To demonstrate the use of regression splines, we have shown in Figure 4 a spline curve to the relationship between the (natural) logarithm of demand and current temperature for all time periods.

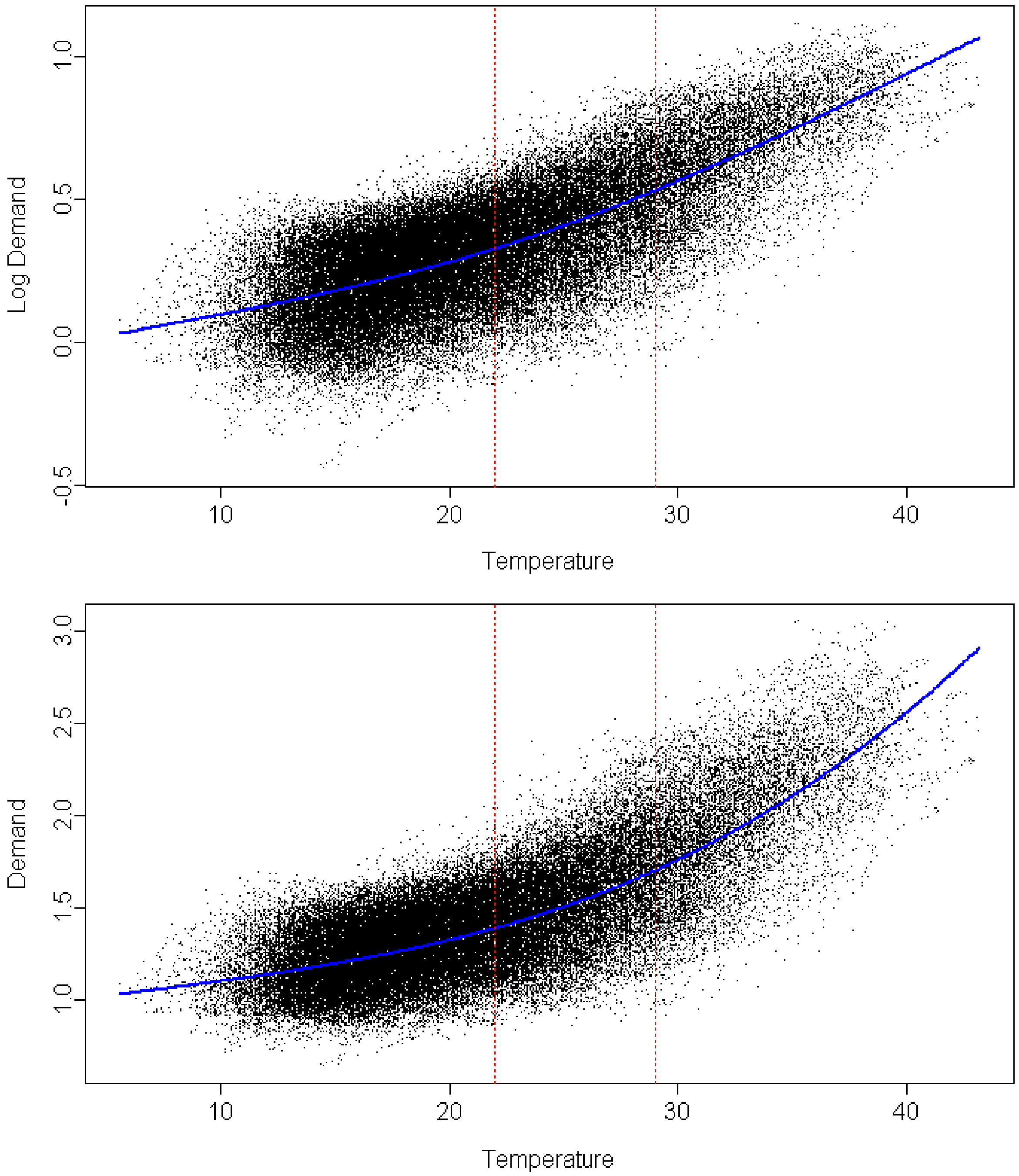

Figure 4: *Example of a cubic regression spline with knots at 22 and 29°C. The curve in each segment in the top panel is a cubic polynomial. The three curves are constrained to ensure the result is continuous and smooth. The bottom panel shows the same curve on the original scale (after taking anti-logarithms).*

A regression spline consists of several polynomial curves that are joined at points known as "knots". In Figure 4, we have used knots at 22 and 29°C (shown by the vertical red dashed lines). The blue curve in each segment in the top panel is a cubic polynomial. The three curves are constrained to ensure the result is continuous and smooth at the knots. The curves are also constrained to give linear relationships at the ends of the curve. The bottom panel shows the same curve on the original scale (after taking anti-logarithms). The choice of knots can be made automatically as explained later in this document.

2.5 Variable selection

The terms in the model are selected by considering their out-of-sample forecasting performance on a withheld sample of data. Provided there is sufficient data, this provides a robust method of model selection that allows a model with good predictive capacity to be chosen.

The out-of-sample forecasting performance is measured using the root mean squared error (RMSE) as it is easy to compute and prefers models without outliers. However, the MAE could also be used.

When there is insufficient data to allow out-of-sample forecasting performance to used in model selection (e.g., with annual data), then the corrected Akaike's Information Criterion (Harrell, 2001, p.202) can be used to select the best model. The corrected AIC can be expressed as a penalized likelihood

$$\text{AIC}_C = L - 2p\left[1 + \frac{p+1}{n-p-1}\right]$$

where L is twice the logarithm of the likelihood of the model, p is the number of parameters in the model and n is the number of observations used in fitting the model. In this formulation, the higher the value of the AIC_C, the better the model.

It can be shown that L is asymptotically equivalent to minus the sample size times the log of the in-sample one-step RMSE plus a constant. So maximizing the AIC_C is similar to minimizing the RMSE.

Whether we use an out-of-sample RMSE or the in-sample AIC_C, the variables to be included as linear functions can be constrained so that their coefficients are of the same sign as expected. For example, because electricity demand is expected to decrease as price increases, the price variable can be excluded if the estimated coefficient of price is positive.

When the variables to be included are uncorrelated with each other, then the coefficients can be interpreted as the effect of the associated variable on the response variable. For example, if price is measured in dollars and demand is measured in MW, then the coefficient of price is the number of MW that demand will increase with an increase of $1 in the price.

However, when there are correlations between the input variables, this interpretation breaks down. Instead, each correlation is the effect of the variable *after taking out the effect of all the other input variables*. When the variables are highly correlated, the effects of the other input variables may largely account for the effect of the input variable of interest. In these circumstances, it is possible for the sign of the estimated coefficient to be opposite to expectation.

If this occurs, it is sometimes argued that it is better to eliminate the variable from consideration (e.g., Harrell, 2001, p.60) as it leads to models that make no intuitive sense. For example, with electricity demand, it may be possible to identify a model with a positive coefficient on price because the increasing price has been included by proxy in the model through another variable (e.g., GSP). However, in these circumstances it is unlikely that the inclusion of price would make much difference to the forecast performance of the model, and it is likely to lead to difficulties in interpretation and in selecting sensible future values of the input variables.

Some have argued against this position on the grounds that our intuition on the right sign of coefficients can be incorrect, and that the best model should be used for forecasting regardless of its interpretive value. However, in the case of electricity demand, we consider that the variables are sufficiently well understood for the direction of effect to be known in advance and that interpretation remains a valid objective with any forecasting model. Furthermore, with highly collinear variables it is necessary to select future values of the variables that have the same degree of collinearity as has occurred in the past (Myers, 2000, pp.379–381). This is very difficult to do in practice.

So, on balance, we favour the elimination of variables with counter-intuitive sign and we believe this strategy will lead to better forecasts.

3 Model estimation

This model is fitted using the R software environment (Core Team, 2014). The demand and temperature data are available on a half-hourly basis, while the economic and demographic data are only available on a seasonal basis. Consequently, to fit this model we employ the following estimation scheme.

- The mean demand $\bar{y}_i$ in each season is calculated.
- The demand is normalized in each year by dividing by the mean demand for that season to give $y^*_{t,p}$
- Model (1) is estimated from the normalized half-hourly data.
- Model (2) is estimated from the mean seasonal demand data.

3.1 Boosting

To estimate (1), we use a procedure based on gradient boosting (Ben Taieb and Hyndman, 2014; Bühlmann and Yu, 2010), involving the following steps.

1. Fit a linear model (i.e., no splines, everything linear), with all possible predictors.
2. The residuals from the linear model are computed and an additive nonlinear model is fitted with the same predictors, using variable selection to decide which predictors to include. Each term in the additive model involves a regression spline with a single knot. The knot location is chosen by cross-validation.
3. The residuals from the preceding model are computed and an additive nonlinear model is fitted with the same predictors, using variable selection to decide which predictors to include. Each term in the additive model involves a regression spline with a single knot. The knot location is chosen by cross-validation.
4. The residuals from the preceding stage are computed, and a further nonlinear model is fitted against the predictions from the preceding stage.

In each of stages 2–4, a shrinkage factor of 50% has been used. The final predictions are the sum of the predictions from all boosting stages.

It would be possible to add further boosting stages, but with a heavier computational burden.

The final boosting step 4 allows for a "saturation effect" in extreme demand periods. This occurs when all available air conditioning facilities are already in use, so an increase in temperature cannot lead to an increase in demand. There is probably also some price sensitivity during periods of high demand. Hence we expect that the residuals from step 3 are possibly non-linearly related to the predictions from step 3, especially during periods of very high demand. Boosting step 4 corrects this effect if necessary.

3.2 Dynamic linear model

To estimate (2), we take first differences for both independent variables and dependent variable in the seasonal demand model (2), to avoid spurious regression resulting from using ordinary least squares to estimate a regression model involving non-stationary time series (Harris and Sollis, 2003). We evaluate the stationarity of the variables in the seasonal demand model using the ADF test, and we usually find that all the variables appear to be non-stationary. Therefore, we take first differences for both independent variables and dependent variable in the seasonal demand model. Thus, we regress the first difference of per-capita annual mean demand against the first differences of the seasonal driver variables, including the first difference of cooling degree-days for each summer and the first difference of heating degree-days for each winter.

An alternative approach that we have used is to fit a model with the same predictor variables as (2), but using dynamic ordinary least squares (DOLS) to take care of the serial correlation and cointegration.

The price coefficient can be very difficult to estimate, especially when there has not been much price variation during the historical data. It is also an important element of the model for considering the effect of potential future price changes. Further, the price elasticity is known to vary with demand — consumers are less sensitive to price at periods of extreme demand which tend to coincide with periods of extreme temperatures. Consequently, we prefer to fix the price coefficient to an appropriate value, and then estimate the remaining coefficients using a least squares approach.

We typically estimate the model without constraints, shrink the estimated price coefficient to a fraction of its former value (with the shrinkage parameter determined by our previous price elasticity analyses), then re-estimate the remaining coefficients.

3.3 Residual analysis

The residuals e_t obtained from the model will have some remaining serial correlation which needs to be accounted for.

Typical residuals (e_t) from model (1) are shown in Figure 5. The top panel shows the residuals with the median of each 35 day block in red. These medians are plotted in Figure 6 which shows that they are serially correlated. We use an ARMA model to match the dynamics of the series. In this example, we identified an AR(1) with coefficient 0.60 as an appropriate model.

The ACF of the residual series (bottom panel of Figure 5) allows us to visualize the serial correlation in the series. There are two striking features in the graph: (1) the strong daily seasonality that has not been captured by the model; and (2) the long-term serial correlation that has not died away even after 28 days of observation. We need to incorporate these phenomena into the forecast distributions when we resample the residuals.

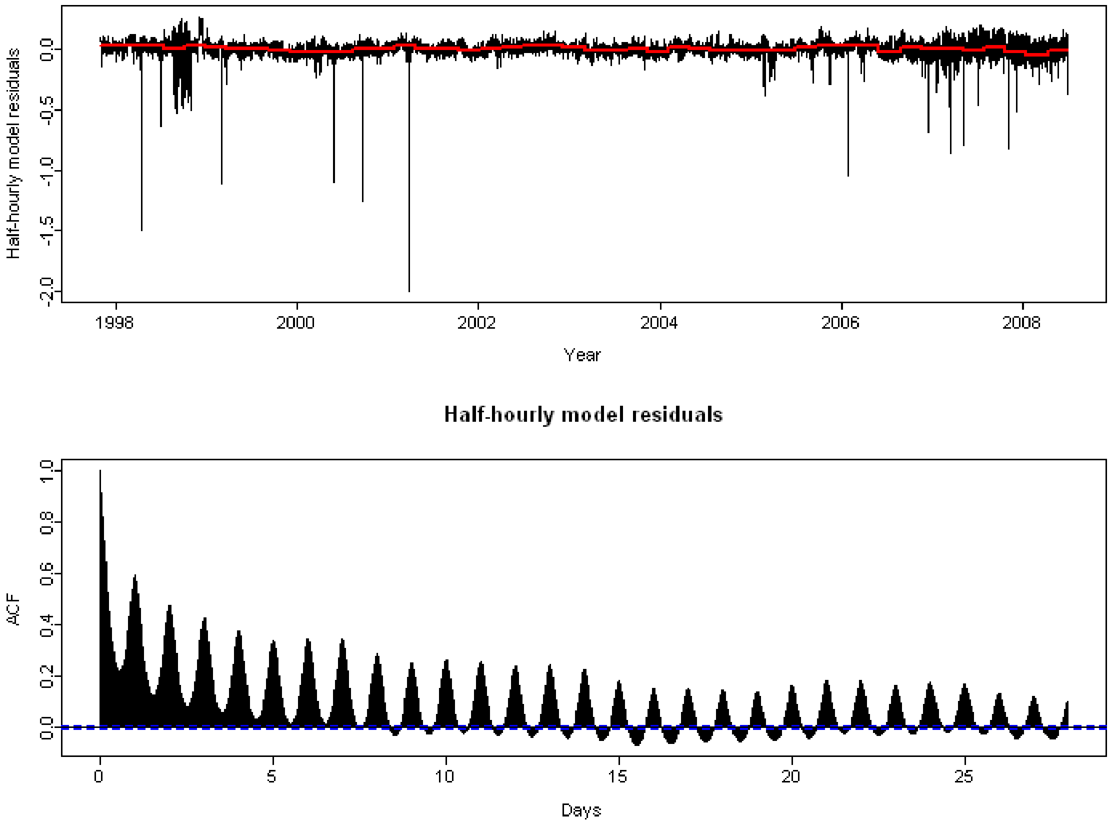

Figure 5: *Top: The residuals from the half-hourly models. The red line shows 35-day medians. Bottom: The ACF of the residuals shown in the top panel.*

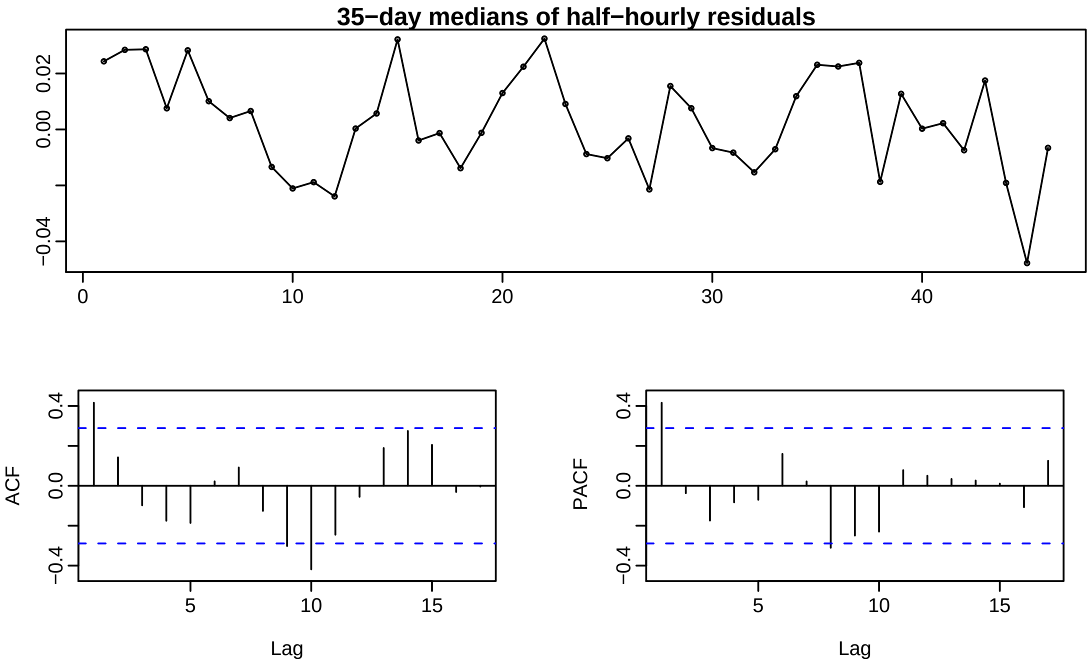

Figure 6: *The series of 35-day medians of a residual series shown in the top panel of Figure 5.*

4 Identifying model changes over time

The model defined by (1) and (2) assumes that the distribution of normalized demand conditional on temperatures and calendar effects does not change over time. That is, all the long-term changes in demand are determined by (2) rather than (1). In particular, the load factor is assumed to be constant, on average, over time. This is likely to be a good approximation over a period of a decade or so, but unlikely to be true over long periods of time.

To check the accuracy of this assumption, we carry out structural break tests on the residual distribution and the load factor.

We compute the residuals from the model and compute the annual 10%, 50% and 90% quantiles. These are plotted over time and the **strucchange** package in R (Zeileis et al., 2002) is used to test for structural breaks in these quantiles of the residual distribution. A similar analysis is used to test for change in the load factor.

We use an empirical fluctuation test based on the OLS-based CUSUM (cumulative sums). This is from the generalized fluctuation test framework, and involves fitting a constant model to the data and deriving an empirical process, which is intended to capture the fluctuations in the residuals. A null hypothesis of “no structural change” is assumed, and this is rejected when the fluctuations of the empirical process become improbably large compared to the fluctuations of the limiting process. Further details are provided in Zeileis et al. (2002). We test the null hypothesis that the model residuals and system load factors remain stationary over the years, by computing the OLS-based CUSUM process and plotting with standard and alternative boundaries.

To illustrate this approach, Figure 7 shows the 10%, 50% and 90% quantiles of the residuals for the NSW summer model, and Figure 8 shows the system load factor in NSW over the same time period. The OLS-based CUSUM processes computed for these time series are given in Figures 9 and 10. Note that none have exceeded their boundaries, indicating that there is no evidence of a structural change in the demand distribution over the years of available data. Empirical fluctuation tests (using function `sctest()`) confirms the above observations.

When working with data from multiple regions, we use a Bonferonni adjustment to take account of the multiple comparisons being made. That is, the significance level is set to 5% when taken across all regions for each of the two structural change tests and for each of the quantiles and load factors. Of course, multiple comparisons are still present because there are several statistics tested within each region, but these are harder to allow for due to the likely correlations between the results. Consequently, we have not attempted to make any adjustments for multiple comparisons within each region.

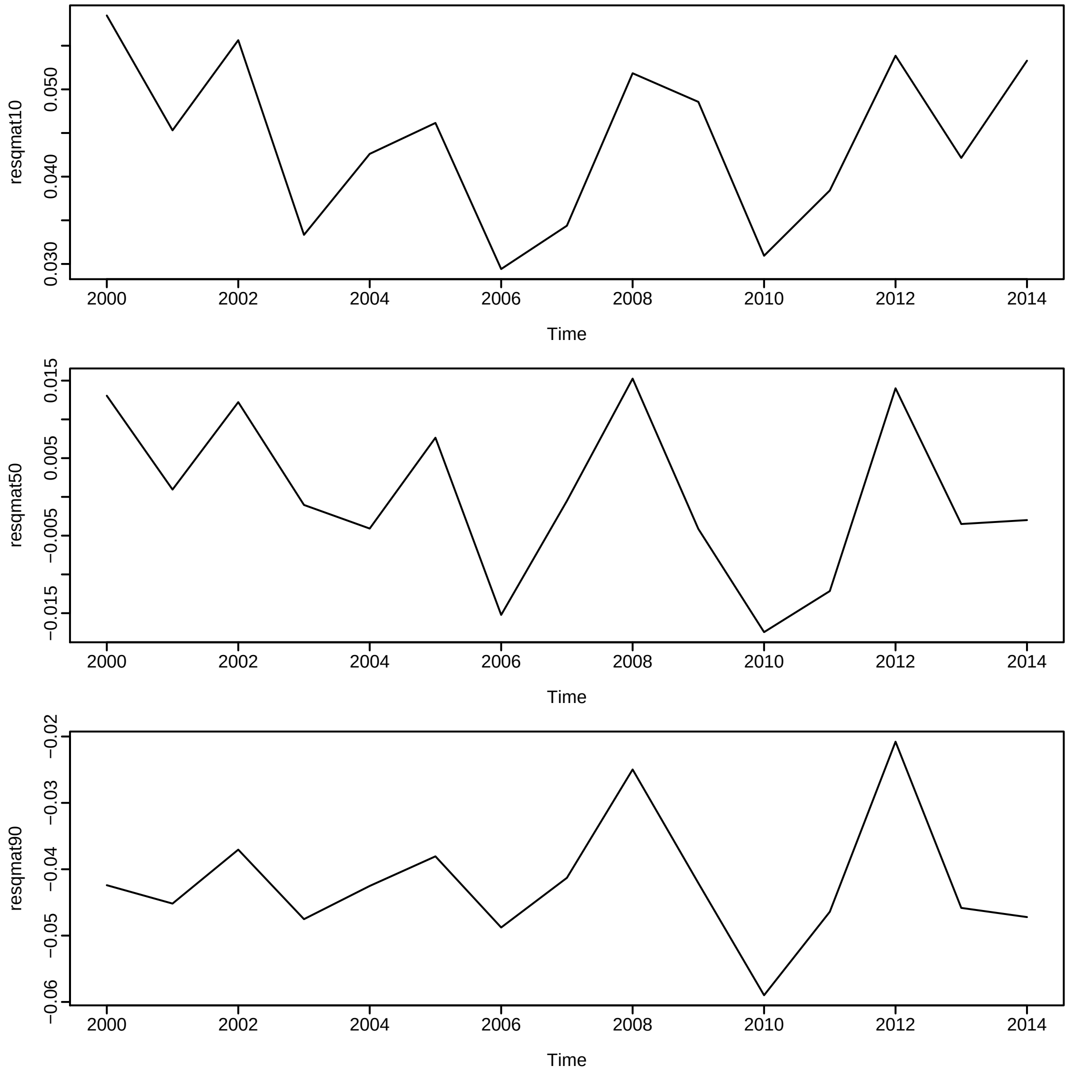

Figure 7: *10%, 50% and 90% residuals quantiles for NSW summer model*

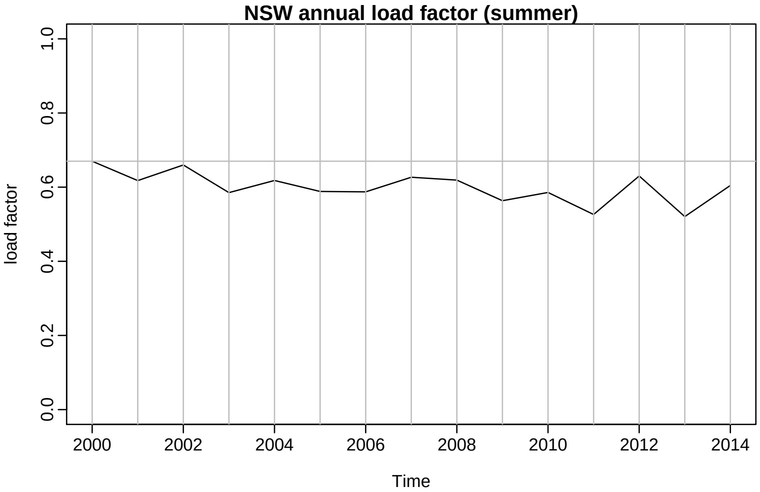

Figure 8: *load factor for NSW summer*

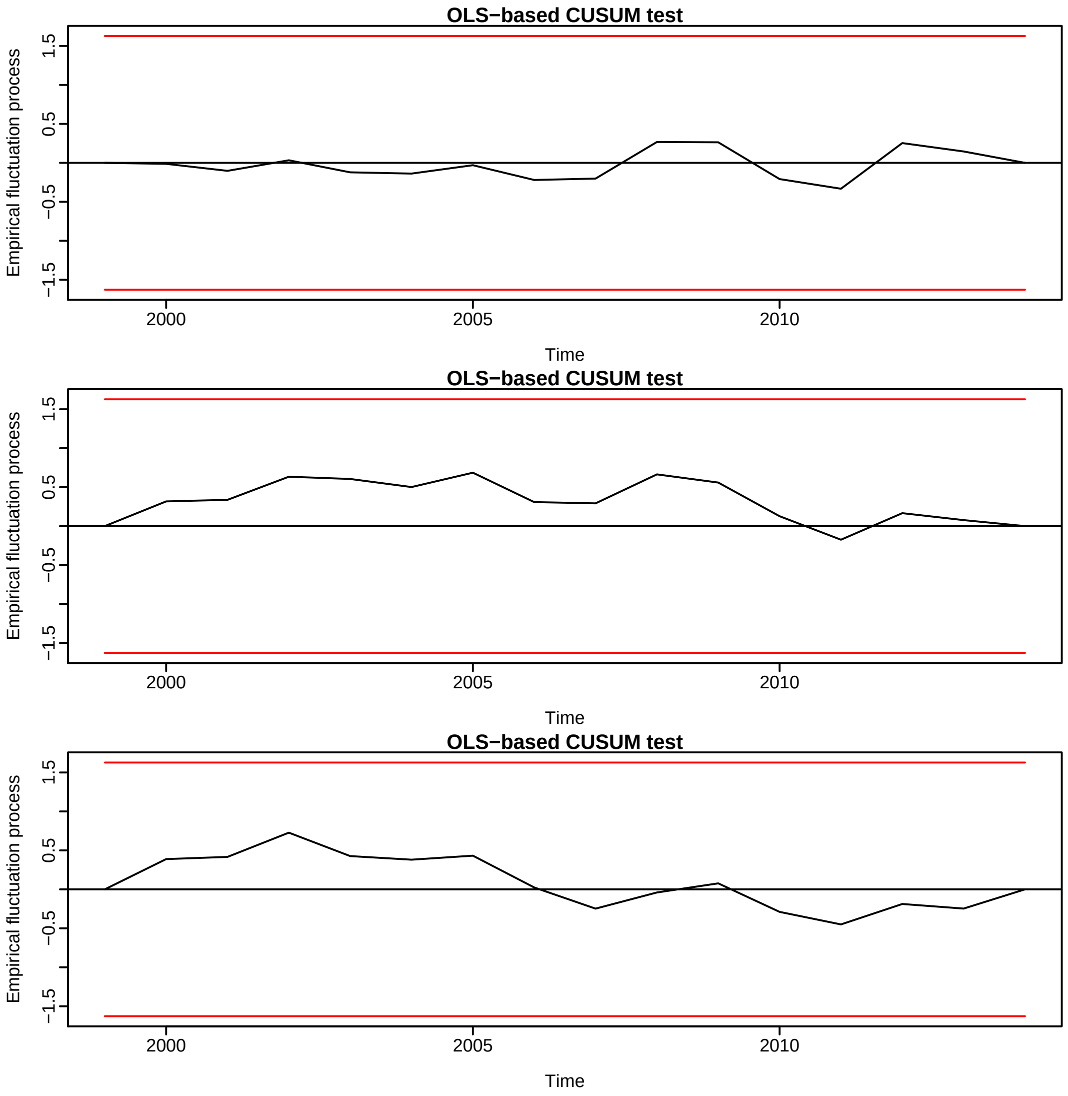

Figure 9: *OLS-based CUSUM processes of 10%, 50% and 90% residuals quantiles and their boundaries for NSW summer model*

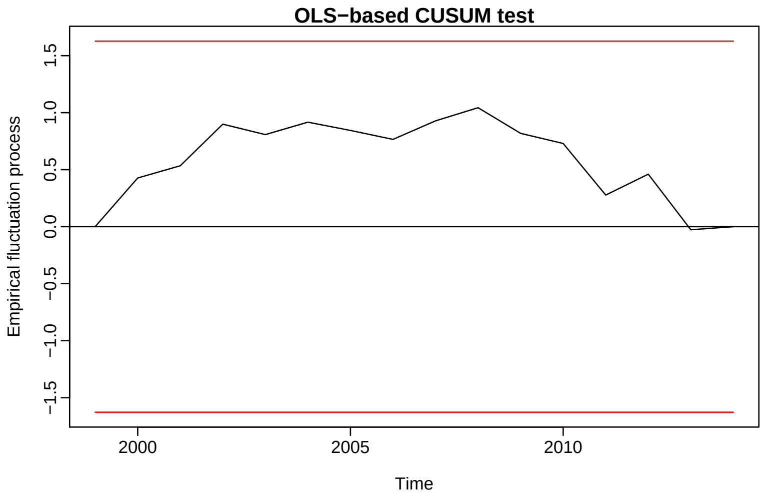

Figure 10: *OLS-based CUSUM processes of the annual load factor and its boundaries for NSW*

5 Model simulation

In order to produce forecasts using our model, we need to simulate future values of temperature and residuals. The future values of other variables in the model are either known in advance (such as calendar variables), or will be assumed to take on some specified values (such as price, economic and demographic variables).

5.1 Temperature simulation

The temperatures are simulated from historical values using a double seasonal bootstrap with variable length as described in Hyndman and Fan (2010). The temperature bootstrap is designed to capture the serial correlation that is present in the data due to weather systems moving across the region.

5.1.1 Half-hourly temperature data

We will use temperature data from Adelaide, South Australia for illustration purposes. Similar results would be obtained if we had used data from the other regions.

Half-hourly temperature data from 1 July 1997 to 12 January 2008 were obtained for Adelaide, Australia. These data were collected from Bureau of Meteorology Adelaide Station. The station is located on the edge of the Adelaide CBD.

Because our interest is in peak load demand, only data from November–March were retained for analysis and modelling. All data from April–October for each year were omitted. Thus, each "year" consists of 182 days. We define the period November–March as "summer" for the purposes of this report. Thus, approximately 10.5 summers of half-hourly data are available.

Each day is divided into 48 periods which correspond with NEM settlement periods. Period 1 is midnight to 0:30am Eastern Standard Time. The half-hourly temperature data come from the Bureau of Meteorology and are the temperatures at the end of each half-hourly period.

Time plots of the half-hourly data are plotted in Figures 11–13. These clearly show the intra-day pattern (of length 48) as well as the annual seasonality (of length $48 \times 182 = 7248$).

Figure 13 shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the half-hourly data. The top plot (ACF) shows the correlation between each period and nearby periods. The horizontal scale is in days. Thus, there is high positive correlation between periods exactly one day apart, and almost no correlation between periods exactly 1.5 days apart.

The bottom plot (PACF) gives the correlation between observations, after allowing for the intervening observations. For example, the PACF at lag 2 is the correlation of temperatures one hour apart after allowing for any correlations that are seen in the observations that are a half-hour apart. In contrast, the ACF at lag 2 is the correlation of temperatures one hour apart with no adjustment made for any

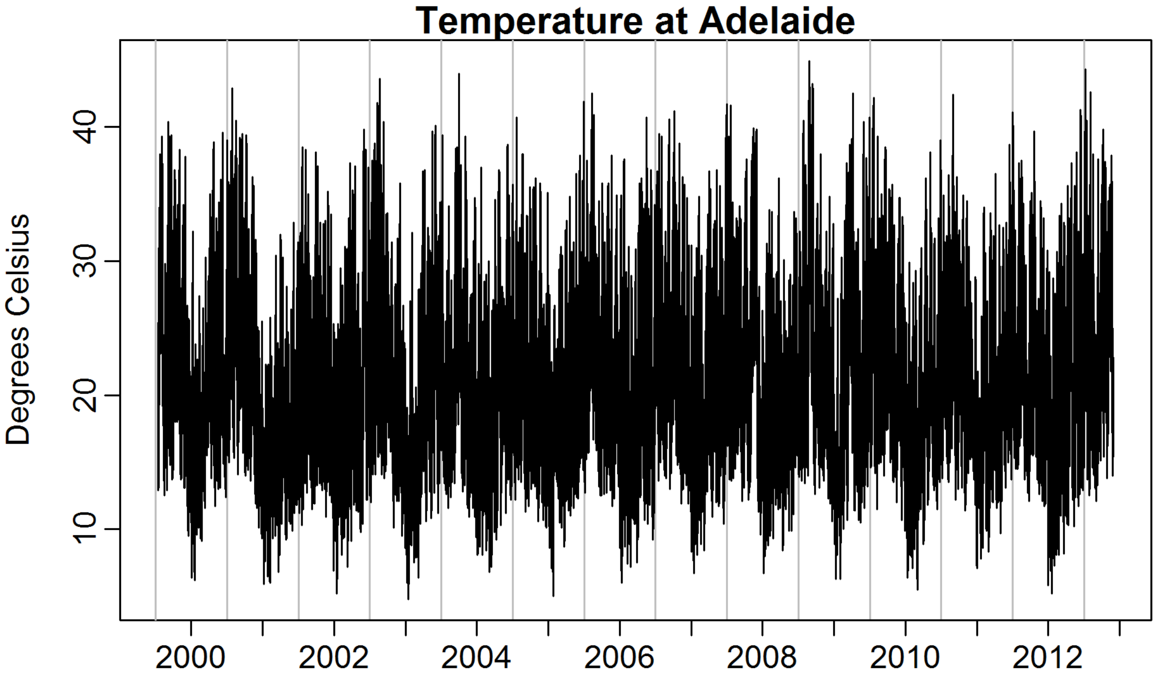

Figure 11: *Half-hourly data at Adelaide from 2000 to 2013.*

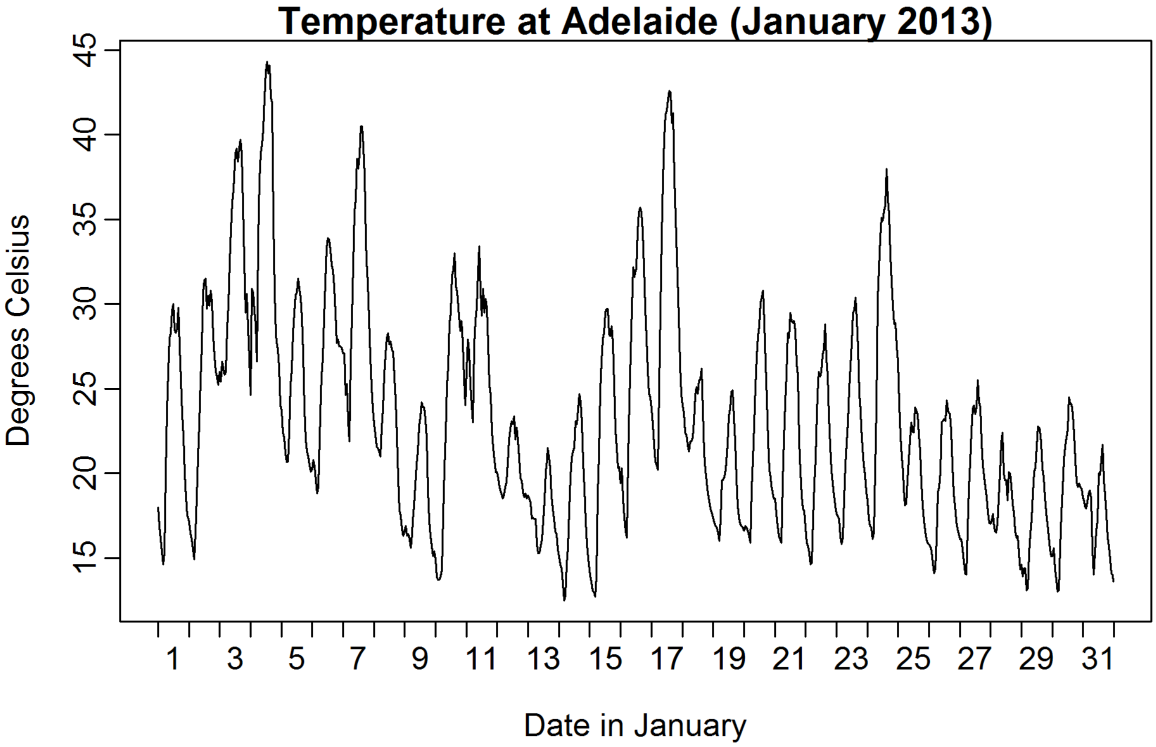

Figure 12: *Half-hourly data at Adelaide from 2000 to 2013.*

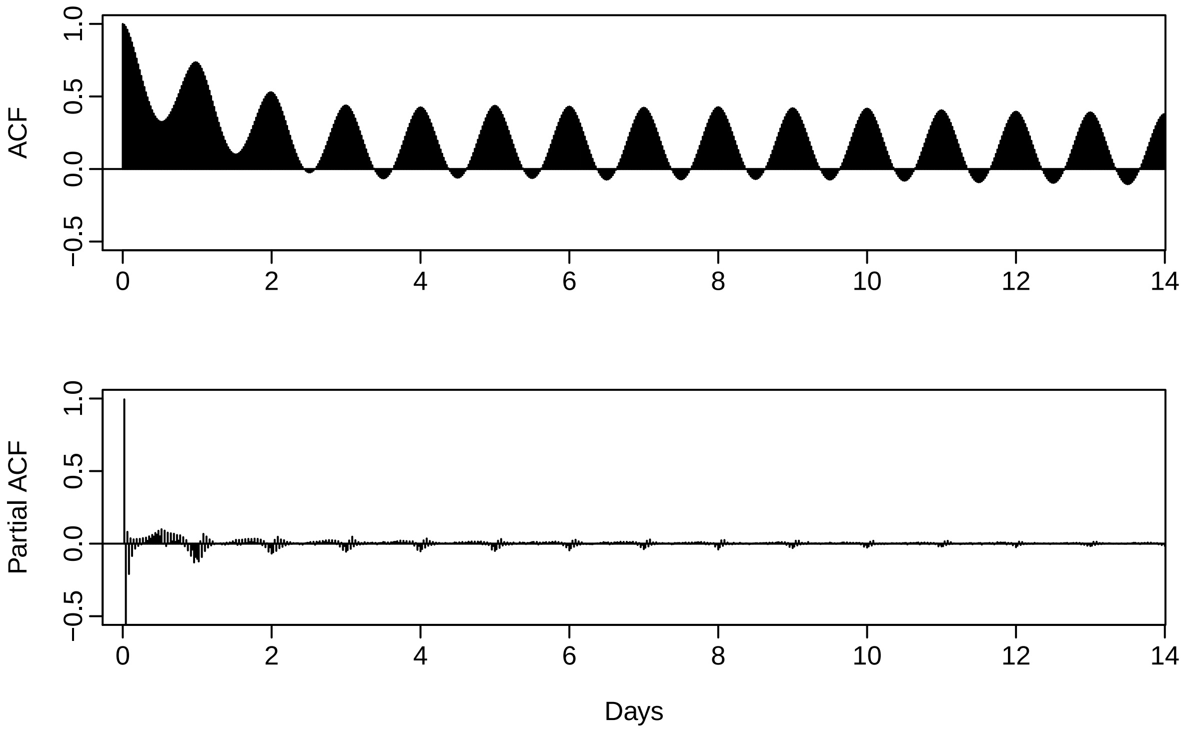

Figure 13: *Autocorrelation and partial autocorrelation functions for half-hourly Adelaide temperatures.*

intervening correlations. The PACF plot shows the daily seasonality persisting for many days, although the correlations are much smaller than in the ACF plot. Both plots are useful in understanding the dynamics of a time series.

The other interesting feature of these plots is that the correlations between periods exactly m days apart continues for $m > 14$. This is known as “long memory” and is probably induced by the annual seasonal pattern in the data, and partly by the time taken for a pressure system to cross South Australia. For example, a high-pressure system in summer will usually take from three to eight days to cross the state, raising temperatures during this period. Consequently, temperatures tend to have relatively long-memory serial correlations. Whatever method of simulation is used, it is important to allow for the long-memory correlation structure as it is directly related to the phenomenon of extended hot periods. If these long-memory correlations are ignored, the simulated temperature data will tend to have only very short hot-spells and therefore electricity peak demand will probably be under-estimated.

Our interest here is in simulating future years of data. It is important that the simulated data represent the true distribution of future temperatures. Thus, our simulated data should match the characteristics shown in all of these plots.

5.1.2 Seasonal bootstrapping

We shall now describe the seasonal bootstrapping methods that will be used for simulation. We assume that the long-term temperature patterns are stable. Of course, with climate change this is untrue, but we will consider how to allow for climate change later.

The method of bootstrapping involves randomly resampling historical data. With time series, it is important to preserve any seasonal or trend patterns as well as the inherent serial correlation. The standard method for bootstrapping time series is the "block bootstrap" Politis, 2003 which involves taking random segments of the historical time series and pasting them together to form new artificial series. There are obviously a very large number of such series that could be formulated in this way. A key parameter in the technique is the length of each segment. This needs to be long enough to capture the essential serial correlations in the data, but short enough to allow a large number of possible simulated series to be generated.

When applied to seasonal time series, it is important that the length of each segment or block is a multiple of the length of the seasonal period. Politis (2001) calls this a "seasonal block bootstrap" although we will call it a "single season block bootstrap" to distinguish it from the double seasonal version to follow.

In a single season block bootstrap, a bootstrap series consists of whole seasons from the historical data, pasted together in a different order from how they were observed. For example, if this was applied to quarterly data from 1980–2007, then the seasonal period is four quarters or one year. A bootstrapped series of the same length as the original series would consist of a shuffled version of the data, where the data for each year stays together but the years are rearranged. For example, it may look like this:

1999	1994	2005	1991	1986	…	1985

In this paper, this approach is applied to residuals from a fitted model which have a seasonal period of 48 half-hourly periods or one day. The series also show some serial correlation up to about 14 days in length. So it is reasonable to sample in blocks of 14 days, or 672 time periods. The whole series would be divided into these blocks of length 14 days. In ten years of data (each of length 182 days), there would be 107 complete blocks as follows:

Block 1	Block 2	Block 3	Block 4	…	Block 107

where Block 1 consists of observations from days 1–14, Block 2 consists of observations from days 15–28, and so on.

Then a bootstrap sample is obtained by simply randomly resampling from the 107 blocks. For example, the bootstrap series may be constructed as follows:

Block 105	Block 73	Block 39	Block 54	…	Block 5

5.1.3 Double season block bootstrap

The above approach is not suitable for temperature simulation because temperatures contain two types of seasonality: daily seasonality as well as annual seasonality. If we were to take the whole year as the seasonal period, there would be too few years to obtain enough variation in the bootstrap samples.

Consequently, we need a new approach. We divide each year of data into blocks of length $48m$ where m is an integer. Thus, each block is of length m days. For the sake of illustration, suppose $m = 9$ days. Then block 1 consists of the first 9 days of the year, block 2 consists of the next 9 days, and so on. There are a total of $182/9 = 16.8$ blocks in each year. The last partial block will consist of only 7 days.

Then the bootstrap series consists of a sample of blocks 1 to 17 where each block may come from a different randomly selected year. For example, block 1 may come from 1999, block 2 from 1996, block 3 from 2003, and so on. The randomly selected years give rise to a large range of possible bootstrapped series.

The idea is best illustrated in a diagram. The original observed data can be arranged as follows.

B1:1997	B2:1997	B3:1997	B4:1997	B5:1997	…	B17:1997
B1:1998	B2:1998	B3:1998	B4:1998	B5:1998	…	B17:1998
B1:1999	B2:1999	B3:1999	B4:1999	B5:1999	…	B17:1999
⋮						
B1:2007	B2:2007	B3:2007	B4:2007	B5:2007	…	B17:2007

Then one possible bootstrap series may look like this.

B1:2007	B2:2005	B3:2002	B4:2001	B5:2006	…	B17:1999
B1:1999	B2:2007	B3:2002	B4:1999	B5:2005	…	B17:1998
B1:2002	B2:1997	B3:2003	B4:2001	B5:2007	…	B17:2004
⋮						
B1:2003	B2:2003	B3:2004	B4:2006	B5:2000	…	B17:1997

The difference between this and the previous single season block bootstrap is that here, the blocks stay at the same time of year as they were observed, although they may randomly move between years.

Figures 14 and 15 show a bootstrapped series obtained using this method. In Figure 15, the top panel shows the actual data from 1997 for the same time of the year. The red lines in Figure 15 show the boundaries of each block. Note that the block from 28 to 36 days consists of the data from the same period in 1997. The other blocks come from other years.

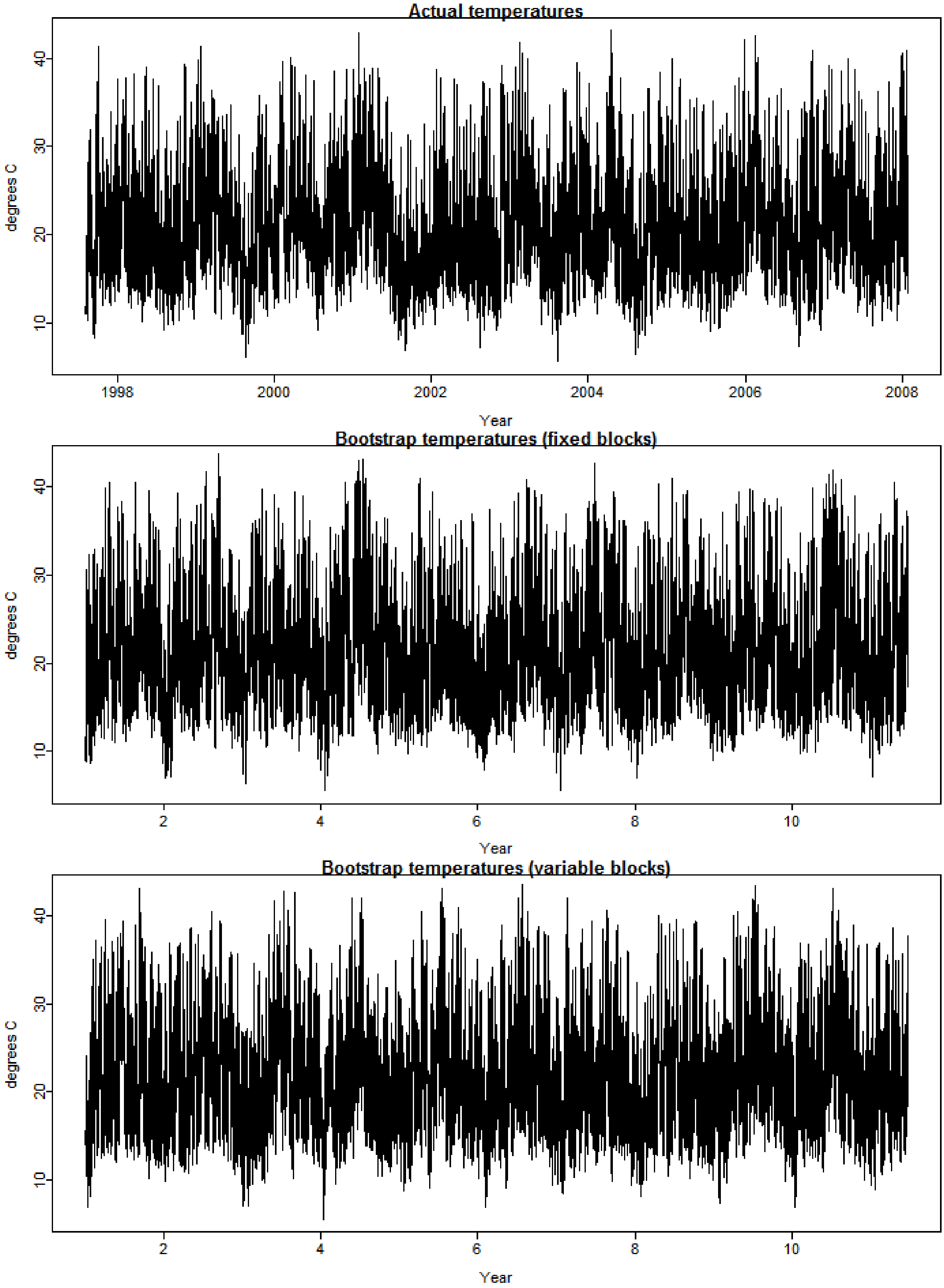

Figure 14: *Actual and simulated temperatures for 10 years.*

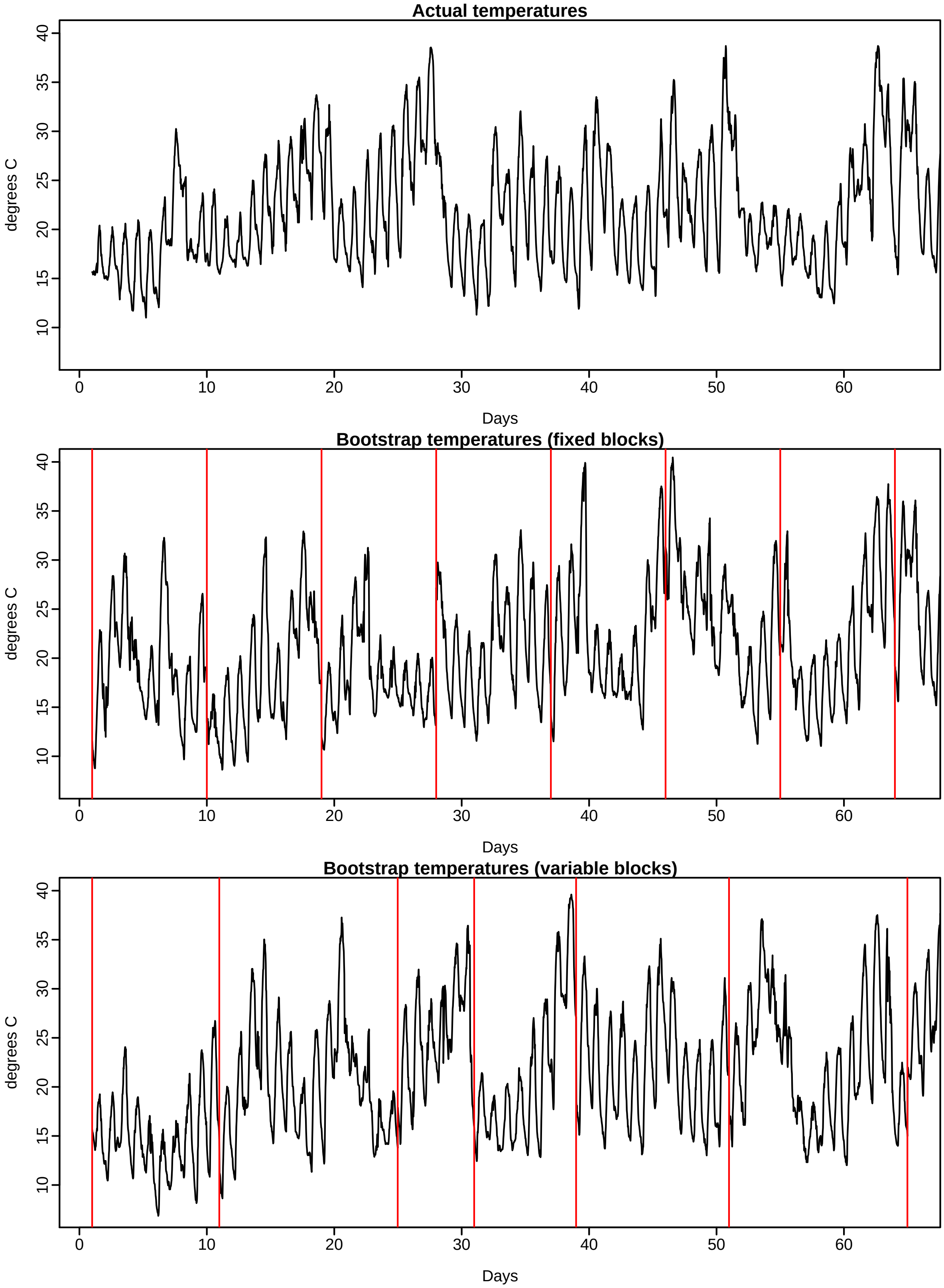

Figure 15: *Top: Actual temperatures from November–December 2000 and some simulated temperatures for November–December showing each block.*

One problem with any block bootstrap method, that is apparent in Figure 15, is that the boundaries between blocks can introduce some large jumps. For example, at the start of day 28, there is a large decrease in temperature from the last period of day 27. This is unrealistic behaviour to be occurring around midnight. However, the behaviour only ever occurs at midnight and we are interested in high temperature behaviour. Thus, the phenomenon is unlikely to be a problem for our simulation purposes.

5.1.4 Double seasonal block bootstrap with variable blocks

As explained above, the number of possible values the simulated series can take on any given day is limited to the specific historical values that have occurred on that day in previous years. With ten years of historical data, there are only ten possible values that any specific point in the simulated series can take. While calculating statistics such as probability levels for individual days, it becomes a problem as there is insufficient variation to obtain accurate estimates.

Consequently, some variation in length and position are introduced. That is, instead of having blocks of fixed length m, we allow the blocks to be of length between $m-\Delta$ days and $m+\Delta$ days where $0 \leq \Delta < m$. Further, instead of requiring blocks to remain in exactly the same location within the year, we allow them to move up to Δ days from their original position. This has little effect on the overall time series patterns provided Δ is relatively small, and allows each temperature in the simulated series to take a greater number of different values.

A uniform distribution is used on $(m-\Delta, m+\Delta)$ to select block length, and an independent uniform distribution on $(-\Delta, \Delta)$ to select the variation on the starting position for each block.

This "double seasonal block bootstrap with variable blocks" performs much better in producing the distribution of temperatures on a given day due to the larger number of potential values the simulated temperatures can take.

In Figure 15, we have used $m = 9$ and $\Delta = 5$. Thus, some blocks are as short as 4 days and others as long as fourteen days. The choice of $\Delta = 5$ also allows temperatures from up to five days earlier or later to be used in the simulation. Figure 15 shows an example where the blocks are of length 10, 14, 6, 8, 12 and 14 days. The starting positions of the blocks varies by 5, 3, 2, 2, -5 and 5 days from their original location. The fifth block in the bottom panel of Figure 15 is from 2000, shown in the top panel of Figure 15, but shifted forward by five days. The first part of this block matches the end of block four in the middle panel.

5.1.5 Adjustments for upper tail

The simulated temperatures are actually adjusted slightly by adding some additional noise to allow the distribution of daily maximum temperatures to closely match those observed since 1900. The hottest day since 1997 has been 44.0°C. To allow hotter days in the bootstrap simulations, we introduce some random noise to the simulated data. This is designed to ensure the density of the bootstrapped series to have the same tail properties as the original data. We cannot add a different noise term to each time period as that gives sample paths that are more 'wiggly' than the original data. Instead, we have a different noise term for each block of bootstrapped data. That is, all time periods within each block receive the same additional noise. The same noise value is added to each time period in a block and to both temperature sites. This ensures the serial correlations and the cross-correlations are preserved.

The added noise term has normal distribution with zero mean and standard deviation

$$\sigma(x) = \begin{cases} 0.3 & \text{if } x \leq 42 \\ 0.3 + 0.4\max(0, x - 42) & \text{otherwise;} \end{cases}$$

where x is the maximum temperature in the block. Thus, a small amount of noise is added to all blocks (to increase the number of possible values that can be obtained in simulations), with additional noise added when the maximum temperature is more than 42. This allows the upper-tail of the distribution to be increased slightly.

5.1.6 Climate change

We can allow for climate change by modifying the noise term that is added to the bootstrap samples to have a positive mean and a larger standard deviation. That is, we replace the noise distribution $\mathrm{N}(0, \sigma(x)^2)$ with $\mathrm{N}(\delta, v(x)^2)$ where $\delta \geq 0$ and $v(x) \geq \sigma(x)$. Although the values of the parameters δ and $v(x)$ cannot be known, some sensitivity analysis can be undertaken by varying these parameters.

To allow for future climate change, some simple climate change adjustments are made to allow for the possible effects of global warming. Estimates were taken from CSIRO modelling (Climate Change in Australia). The shifts in temperature till 2030 for mainland Australia are predicted to be 0.6°C, 0.9°C and 1.3°C for the 10th percentile, 50th percentile and 90th percentile respectively, and the shifts in temperature till 2030 for Tasmania are predicted to be 0.5°C, 0.8°C and 1.0°C for the 10th percentile, 50th percentile and 90th percentile respectively (CSIRO, 2015). The temperature projections are given relative to the period 1986–2005 (referred to as the 1995 baseline for convenience). These shifts are implemented linearly from 2015 to 2035. CSIRO predicts that the change in the standard deviation of temperature will be minimal in Australia. This matches work by (Räisänen, 2002).

Climate change is likely to have much more complex and subtle effects on future temperature patterns and distributions, but this simple scheme may at least give some way of exploring the effect of increasing temperatures.

5.2 Residual simulation

To simulate a set of residuals, first the medians are subtracted from the half-hourly residuals and the 'centered' data are simulated using a single seasonal bootstrap of length 35 days. Then simulated 35-day averages are obtained using an AR(1) model with normal errors. These simulated averages are added into the centred residuals obtained from the seasonal bootstrap to give the final set of simulated residuals.

To test this procedure, Figure 16 shows some original half-hourly residuals from our model, their ACF, along with the ACF of some simulated residuals. This provides quite a close match to the ACF of the actual residuals in the middle panel. We have tried many variations on this procedure in an attempt to get an ACF with as close a match as possible, and the above method is the best obtained.

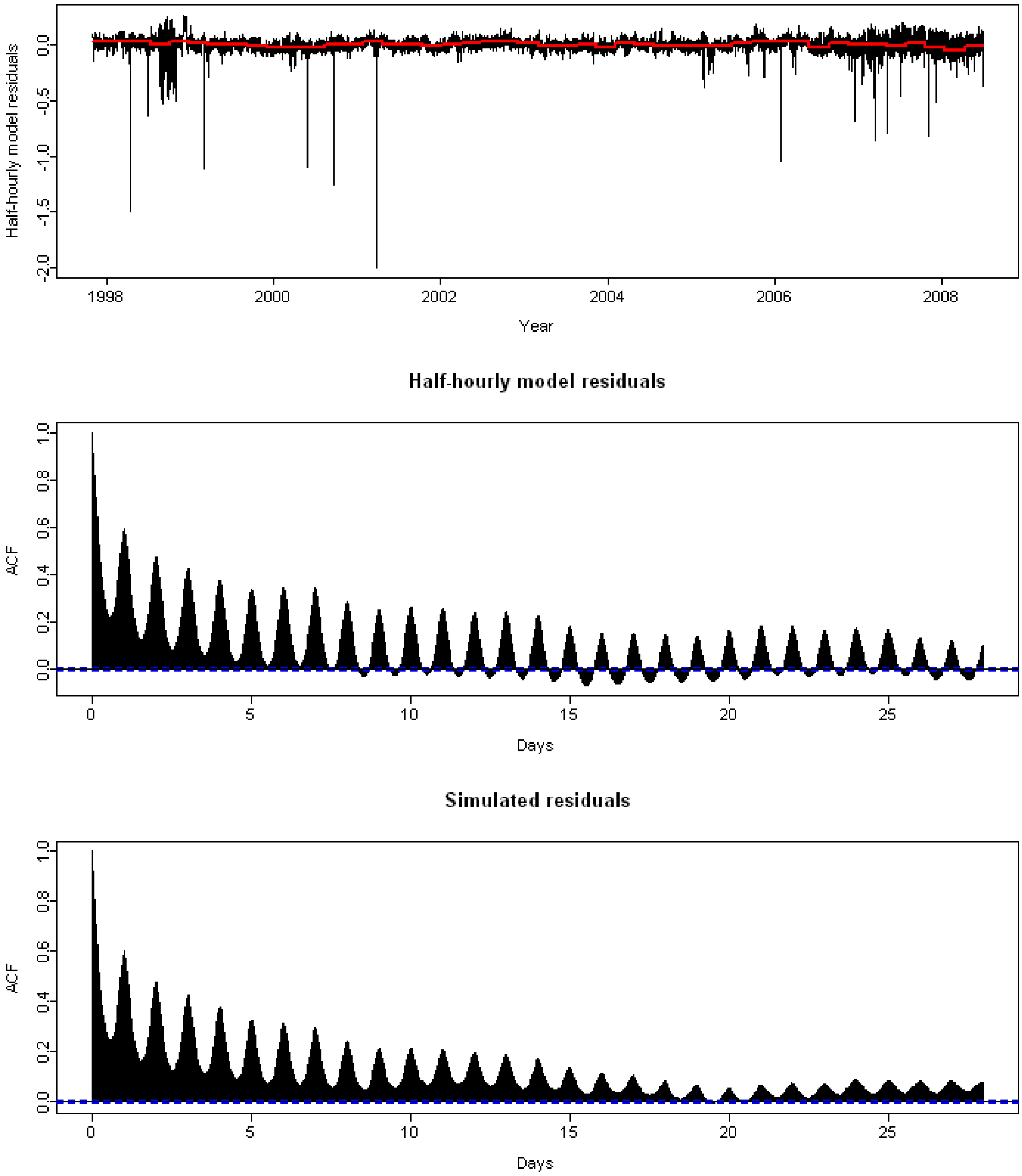

Figure 16: *Top: Actual residuals from a half-hourly model; Middle: The ACF of the residuals in the top panel; Bottom: The ACF of some simulated residuals.*

6 Modelling and simulation of PV generation

Many users of electricity are now generating some of their own power via small-scale rooftop installations of photo-voltaic (PV) cells. Our model is designed to capture the total demand, yet any demand that is met using PV generation is missing from our data.

Consequently, where necessary we use a separate model of PV generation. This is used to augment the original data so that we have estimated demand as if PV generated demand was observed. It is also used to simulate future PV generation, in order to subtract its effect from the simulated future demand, in order to better understand the true generation needs for the future.

Rooftop PV generation is difficult to simulate for future time periods because we have only daily observations of solar radiation, but we want half-hourly forecasts of PV generation. Our approach has been to model the daily PV generation as a nonlinear, nonparametric function of daily solar radiation, maximum temperature and the day of the season.

Then the half-hourly PV generation is normalized by the corresponding daily totals, and simulated with a double seasonal bootstrap using variable length blocks, in the same way as we simulate temperatures and residuals. We also jointly simulate daily solar radiation and maximum temperatures using a single seasonal bootstrap with variable block length. These values are then used in the model to simulate daily PV generation (adding in the model residuals to retain the appropriate level of variability). Finally, the simulated normalized half-hourly PV generated values are multiplied by the simulated daily generation PV values and the seasonal projection of PV installed capacity in the future.

Note that the solar radiation and temperature simulations will be done simultaneously in the forecasting phase of the MEFM in order to preserve correlations between meteorological variables.

We take NSW as an example to illustrate the modelling of PV generation. Figure 17 shows the relationship between the daily PV generation and the daily solar exposure in NSW from 2003 to 2011, and the strong correlation between the two variables is evident. Next we plot the daily PV generation against daily maximum temperature in NSW for the same period in Figure 18, and we can observe the positive correlation between the two variables. Accordingly, the daily PV exposure and maximum temperature will be considered in the PV generation model.

The model for the daily solar PV generation can be written as

$$\log(y_t) = h_t + w_t + n_t \tag{7}$$

where

- y_t denotes the solar PV generation at time t (here measured in daily intervals) ;

- h_t models calendar effect, day of the season is used here;
- w_t models the weather effects, including solar exposure and maximum temperature;
- n_t denotes the model error at time t.

The input variables of the daily PV generation model are selected by minimizing the out-of-sample forecasting errors.

The model fitted results are plotted in Figure 19, and the model residuals are shown in Figure 20.

While it is not possible to validate the forecast accuracy of our approach (as there are no actual half-hourly data to validate against), the above analysis has shown that it is possible to simulate future half-hourly PV generation in a manner that is consistent with both the available historical data and the future temperature simulations.

To illustrate the simulated half-hourly PV generation, we plot the boxplot of simulated PV output based on a 1 MW solar system in Figure 21, while the boxplot of the historical ROAM data based on a 1 MW solar system is shown in Figure 22. Comparing the two figures, we can see that the main features of the two data sets are generally consistent. Some more extreme values are seen in the simulated data set — these arise because there are many more observations in the simulated data set, so the probability of extremes occurring somewhere in the data is much higher. However, the quantiles are very similar in both plots showing that the underlying distributions are similar.

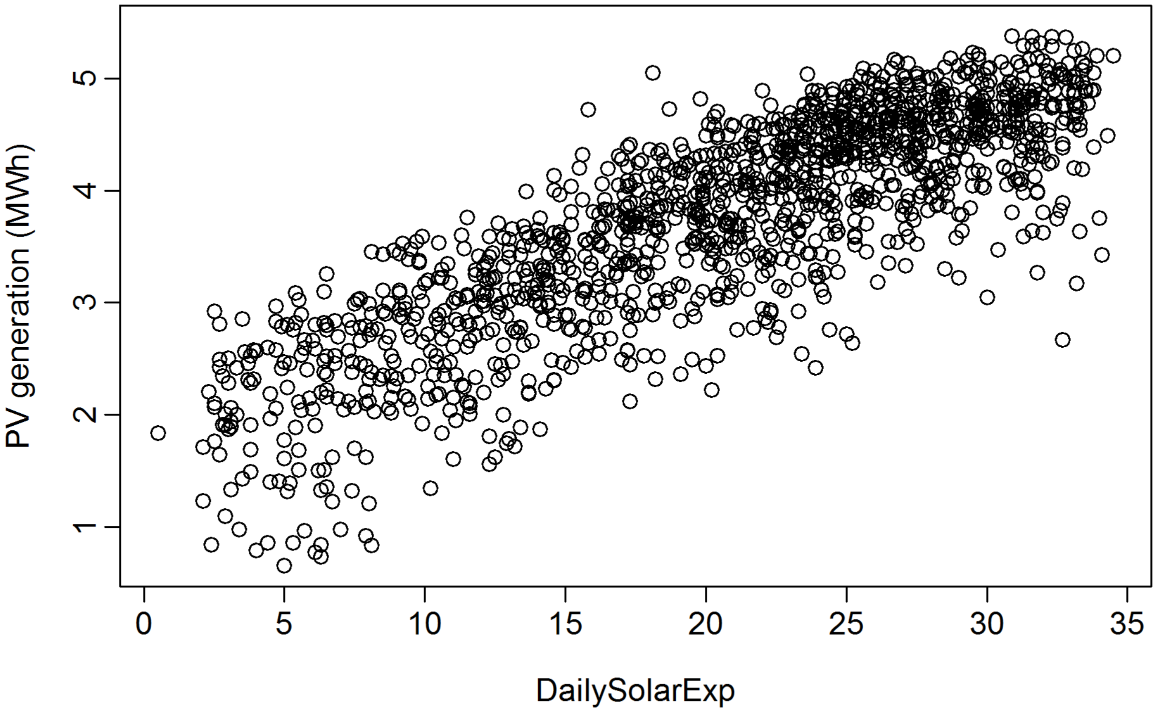

Figure 17: *Daily solar PV generation plotted against daily solar exposure data in NSW from 2003 to 2011. Only data from October–March are shown.*

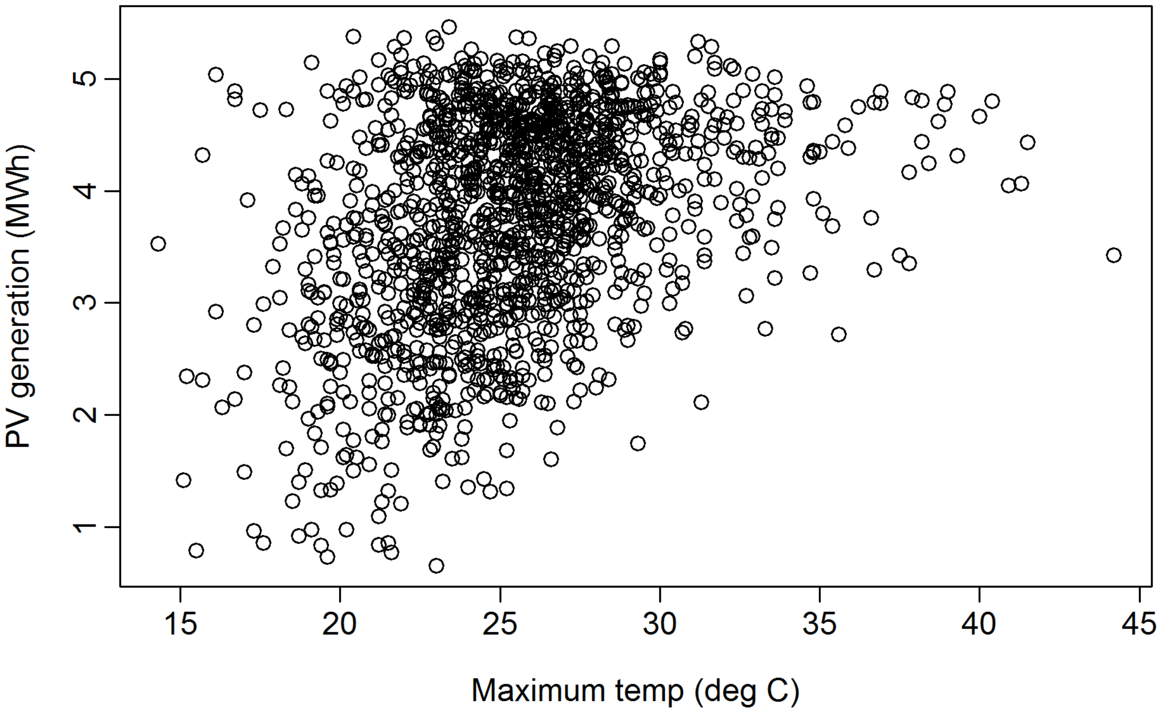

Figure 18: *Daily solar PV generation plotted against daily maximum temperature in NSW from 2003 to 2011. Only data from October–March are shown.*

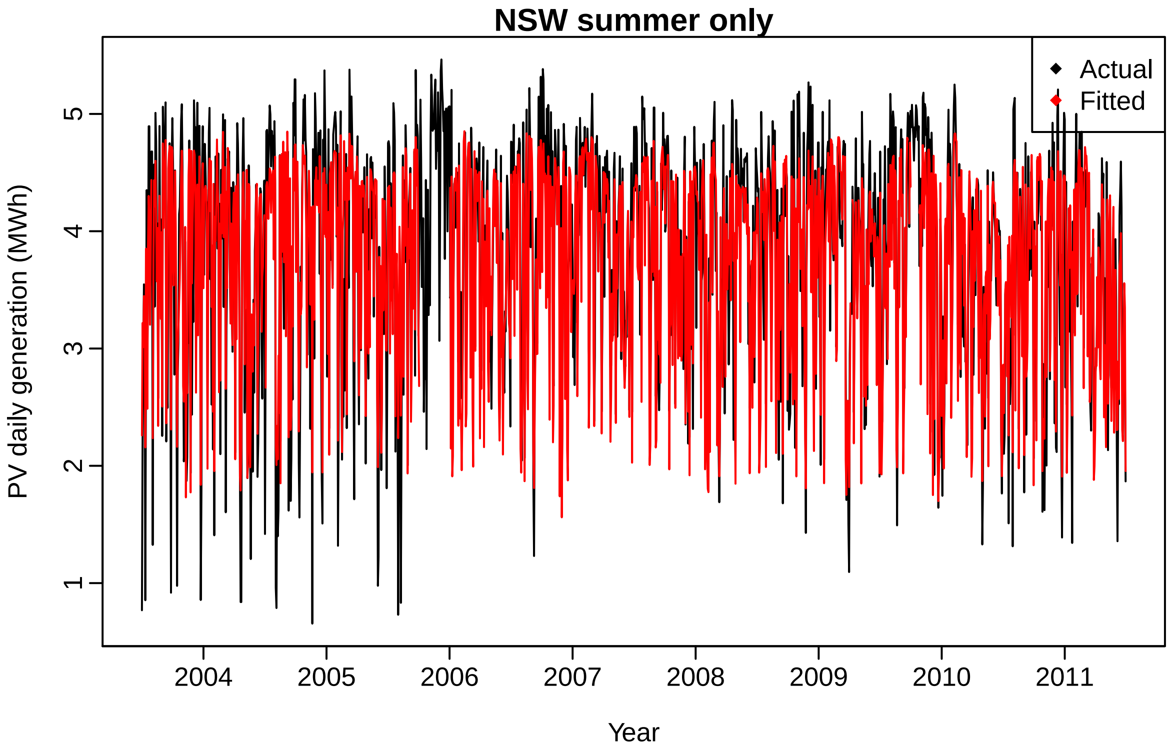

Figure 19: *Actual and fitted PV generation for NSW from 2003 to 2011. Only data from October–March are shown.*

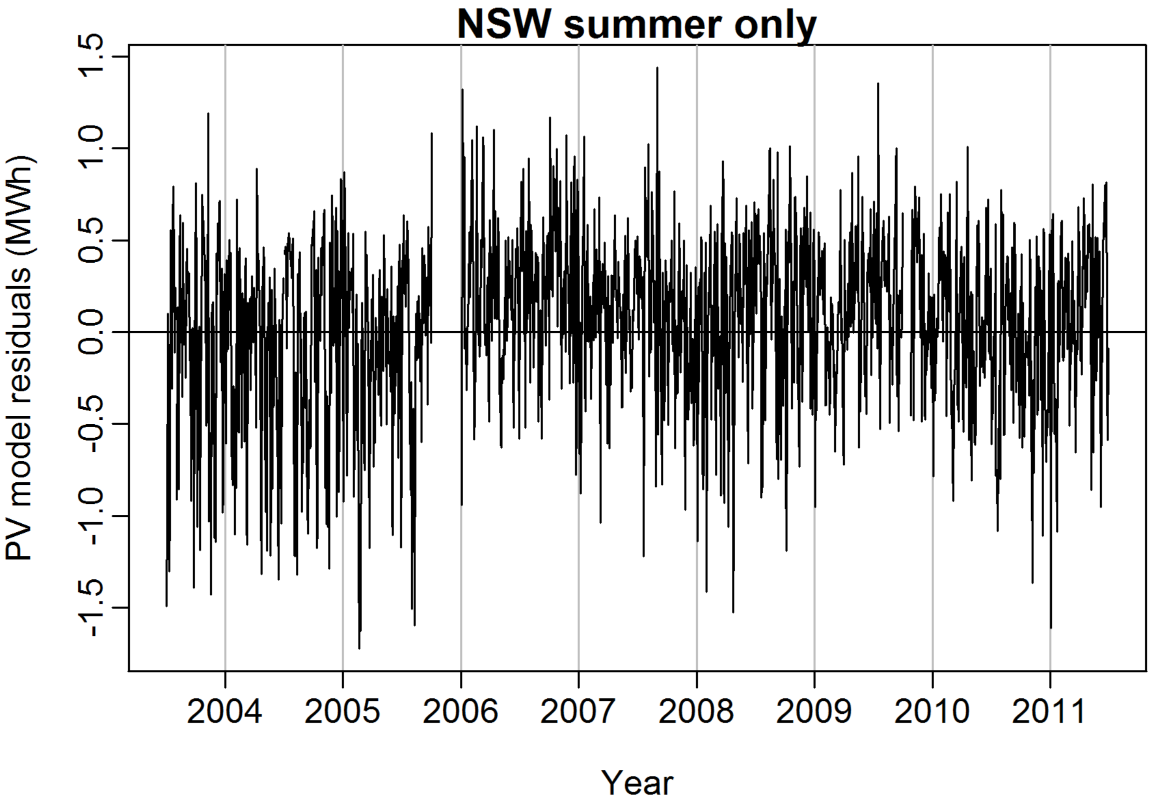

Figure 20: *PV model residuals for NSW from 2003 to 2011. Only data from October–March are shown.*

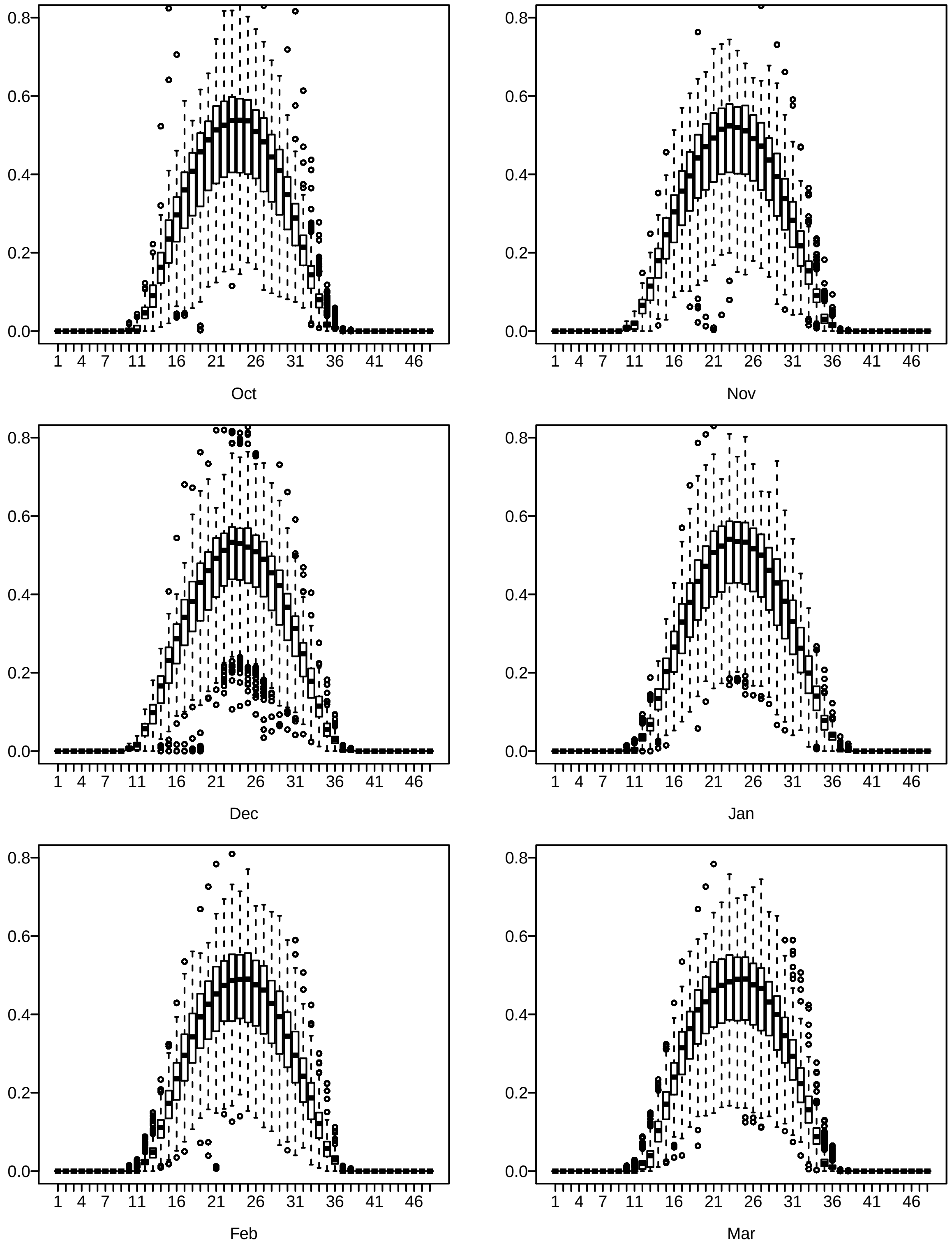

Figure 21: *Boxplot of simulated PV output based on a 1 MW solar system. Only data from October–March are shown.*

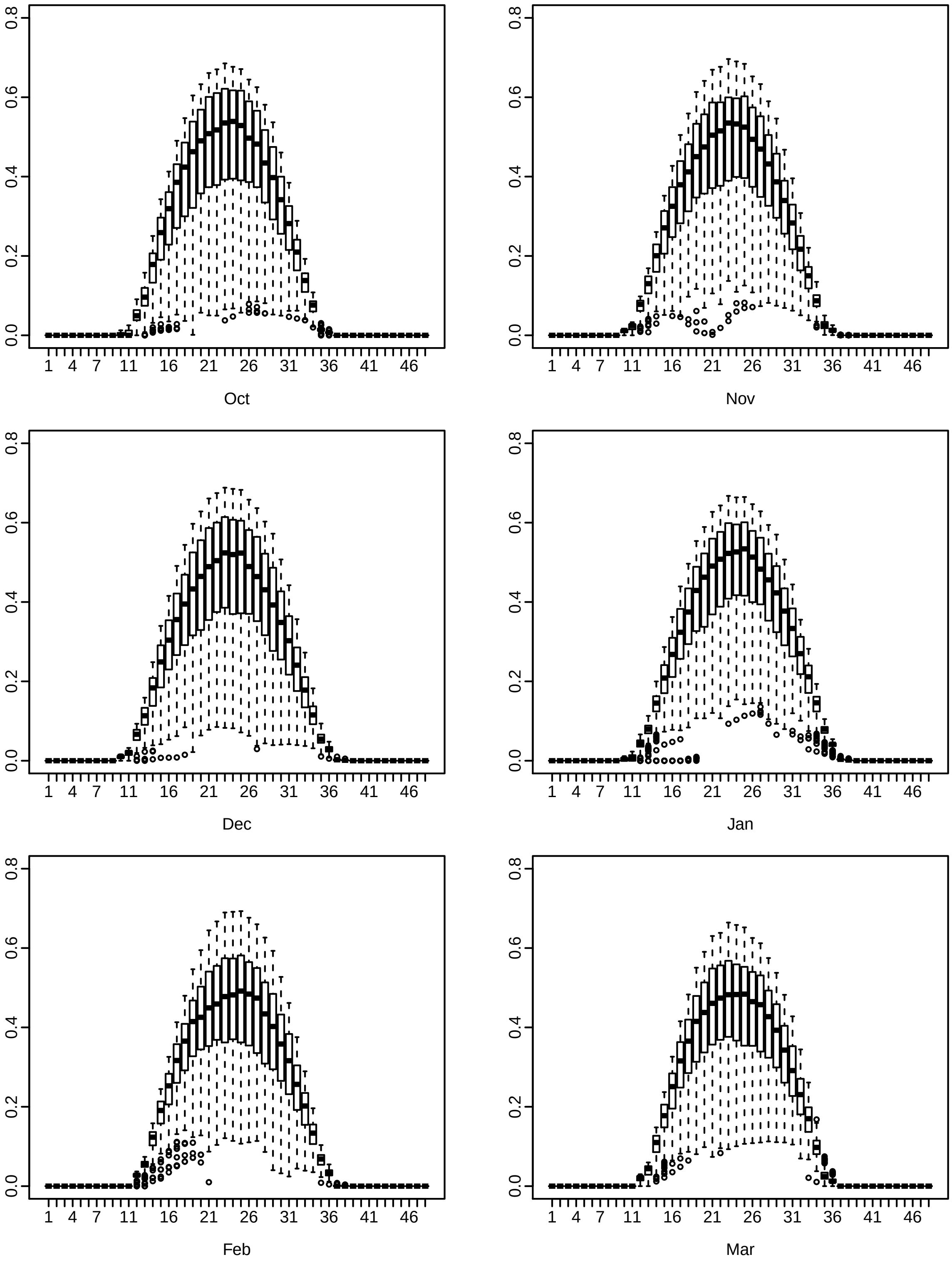

Figure 22: *Boxplot of historical PV output based on a 1 MW solar system. Only data from October–March are shown.*

7 Forecasting

Forecasts of the distribution of demand are computed by simulation from the fitted model as described in Section 5. Specified future values of the demographic, economic and price variables are used, and future temperatures and residuals are simulated. In this way, we generate about 1000 possible random futures consisting of half-hourly demand for all years in the forecast period.

Thus, we are predicting what could happen in the future years under fixed economic and demographic conditions, but allowing for random variation in temperature events and other conditions.

For each future year we can compute a range of possible statistics. The PoE values of the simulated future data are computed using sample quantiles applied to the annual maximums within the simulated data. That is, for each of the 1000 possible futures, we find the annual maximum in each year. Then the 90th percentile (for example) of the 1000 annual maximums in a given year is equal to the 10% PoE value.

7.1 Forecast evaluation

To evaluate the forecasting performance of the model, we compare the actual demand of the historical data, with two different type of predictions: ex ante forecasts and ex post forecasts.

Specifically, **ex ante forecasts** are those that are made using only the information that is available in advance. If we have data to time T, then we calculate future demand using economic conditions as assumed at time T and simulated temperatures based on models for data up to time T.

On the other hand, **ex post forecasts** are those that are made using known information on the “driver variables”. In this evaluation, ex post forecasts for a summer are calculated using known economic conditions and known temperatures for the period being forecast. Note that we do not use demand data from the forecast period for the model estimation or variable selection. Ex post forecasts can assume knowledge of the input or driver variables, but should not assume knowledge of the data that are to be forecast.

The difference between the ex ante forecasts and ex post forecasts provide a measure of the effectiveness of the model for forecasting (taking out the effect of the forecast errors in the input variables).

To illustrate, Figure 23 illustrates the ex ante forecast density function for maximum weekly demand and maximum annual demand for South Australia in 2010/11.

It is difficult to evaluate distributional accuracy on annual maximums because we only see one actual value per year. So the weekly maxima are usually more informative for evaluating distributions.

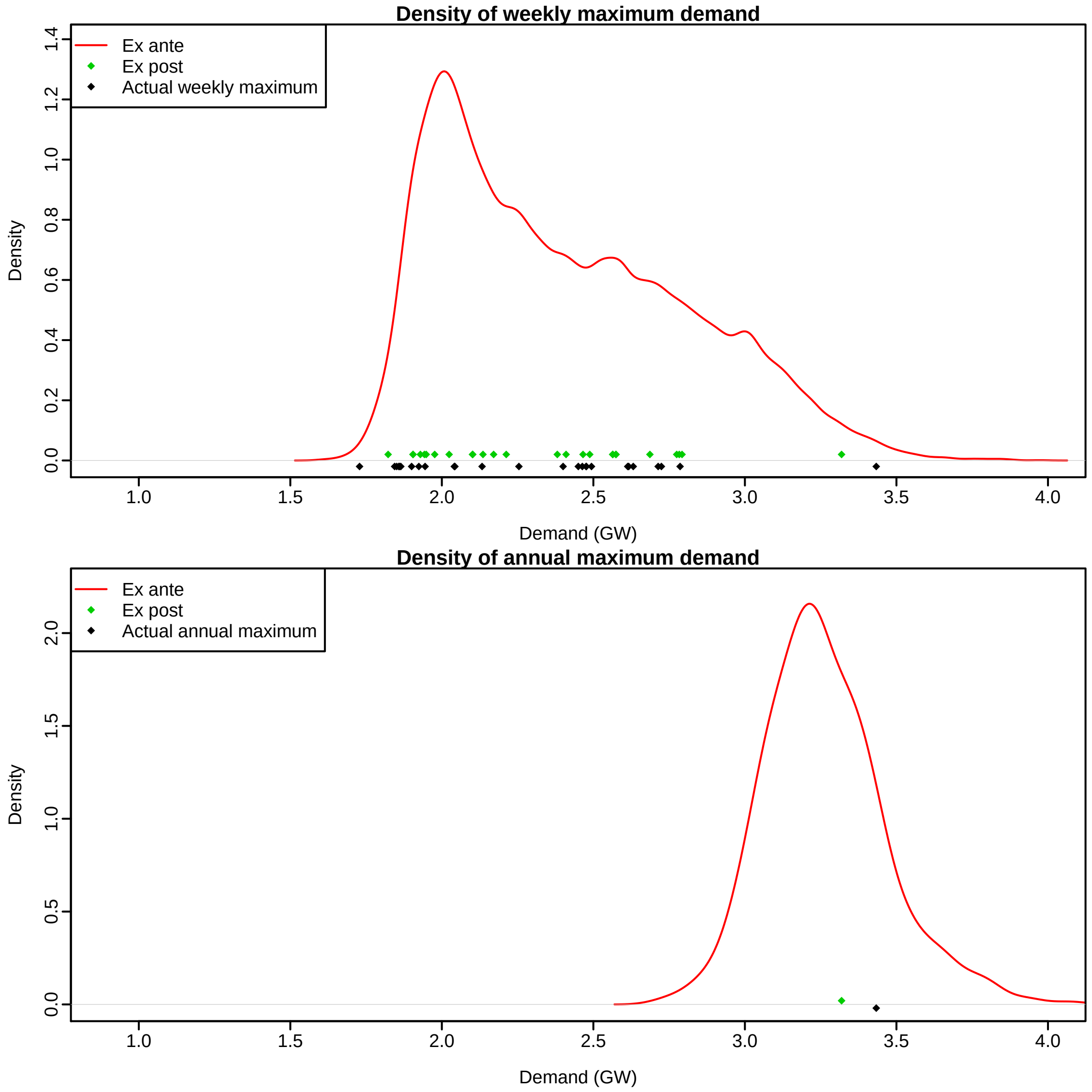

Figure 23: *Ex ante probability density functions for weekly maximum demand and annual maximum demand. Actual values and ex post forecasts are also shown.*

These graphs demonstrate that the actual demand values fit the ex ante forecast distributions remarkably well. In this case, the 26 actual weekly maximum demand values all fall within the region predicted from the ex ante forecast distribution. Although there is only one annual maximum demand observed, the bottom graph shows that this also falls well within the predicted region.

7.2 PoE evaluation and the binomial distribution

A useful way to evaluate the forecast accuracy of PoE values is to consider the percentage of times each PoE value is exceeded in the historical data. For example, we expect 10% of historical maximums to exceed the 10% PoE. If we find that only 8% of values are above the 10% PoE, we can use the binomial distribution to check if that is within the range expected.

More generally, suppose we wish to evaluate the estimated p% PoE values and we observe x out of n observations greater these PoE values. Then using a binomial distribution, the probability of observing at least x values above the PoE values is given by

$$q = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Provided this value is between 5% and 95%, then our forecast distribution can be considered to be representing the historical variation accurately (equivalent to a binomial test with level 10%). Even if the forecast distribution is correct, there is a 10% chance of q being outside the range $(5\%, 95\%)$.

The above binomial test assumes that the maxima are independent. That is true for annual maxima, but is not true for weekly maxima as serial correlation within demand data has longer significant lags than one week. Therefore, this test is most useful on the annual maxima.

References

Ben Taieb, S and R J Hyndman (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting* **30**(2), 382–394.

Box, G E P and D R Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B* **26**(2), 211–252.

Bühlmann, P and B Yu (2010). Boosting. *Wiley Interdiscip. Rev. Comput. Stat.* **2**(1), 69–74.

Core Team, R (2014). *R: A Language and Environment for Statistical Computing*. 09/18/2009. Vienna, Austria, `http://www.r-project.org`.

CSIRO (2015). *Technical Report - Climate Change in Australia - Projections for Australia's NRM Regions*. Accessed: 2015-3-05. `http://www.climatechangeinaustralia.gov.au/en/publications-library/technical-report/`.

Fan, S and R J Hyndman (2010). Short-term load forecasting based on a semi-parametric additive model. In: *Proceedings, 20th Australasian Universities Power Engineering Conference*. University of Canterbury, Christchurch, New Zealand.

Harrell Jr, F E (2001). *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harris, R and R Sollis (2003). *Applied time series modelling and forecasting*. John Wiley & Sons.

Hastie, T J and R Tibshirani (1995). *Generalized additive models*. Chapman & Hall/CRC.

Hyndman, R J and S Fan (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems* **25**(2), 1142–1153.

McSharry, P E, S Bouwman, and G Bloemhof (2005). Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Transactions on Power Systems* **20**, 1166–1172.

Myers, R H (2000). *Classical and modern regression with applications*. Duxbury.

Politis, D N (2001). Resampling time series with seasonal components. In: *Frontiers in data mining and bioinformatics: Proceedings of the 33rd symposium on the interface of computing science and statistics*, pp.619–621.

Politis, D N (2003). The impact of bootstrap mathods on time series analysis. *Statistical Science* **18**(2), 219–230.

Räisänen, J (2002). CO_2-Induced Changes in Interannual Temperature and Precipitation Variability in 19 CMIP2 Experiments. *Journal of Climate* **15**(17), 2395–2411.

Ramanathan, R, R F Engle, C W J Granger, F Vahid, and C Brace (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting* **13**, 161–174.

Ruppert, D, M P Wand, and R J Carroll (2003). *Semiparametric regression*. Cambridge University Press.

Zeileis, A, F Leisch, K Hornik, and C Kleiber (2002). strucchange. An R package for testing for structural change in linear regression models. *Journal of Statistical Software* **7**(2).