Shining Through: Net Load Forecasting in Solar-Rich South Australia

Chase Robertson

School of Computer Science
The University of Auckland

Supervisor: Simon Urbanek

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Data Science, The University of Auckland, 2023.

Abstract

Short-Term Load Forecasting (STLF) is the task of predicting consumers' electricity demand within the next 24 to 48 hours. This task has historically depended primarily on forecasting consumer behaviour. However, with increasing climate and energy-independence concerns, and the rapid growth of rooftop solar panel installation packages, programs, and incentives, the forecasting task now also depends on the productivity of solar and other behind-the-meter generation systems. This issue is examined in the context of two states in Australia: New South Wales, where behind-the-meter generation is a lesser factor in prediction, and South Australia, where behind-the-meter generation has a significant impact on net demand. The regional difference in generation patterns is found to be reflected in forecasting accuracy, with new meteorological inputs like irradiation improving performance significantly. Generalised Additive Mixed Model, Random Forest, and Histogram Gradient Boosting methods are applied, and the latter ensemble methods are found to predict most accurately. Exogenous social information and methodological choices, e.g. holiday encodings and training data windowing strategies, are found to have some impact on performance as well. Our proposed architecture, Histogram Gradient Boosting models trained on annual data windows with a binary-encoded holiday covariate, is found to achieve 65 MAE, 6% MAPE, and 84 RMSE.

Contents

Chapter 1

Introduction

Short-Term Load Forecasting (STLF) is the task of predicting net demand of businesses and consumers on their electricity grid within the next 24 to 48 hours. Grid operators and electricity providers find these forecasts indispensable for providing sufficient power to their customers, while avoiding the waste of overproduction. The financial stakes of overproduction are high, as energy prices fluctuate with the market and geopolitical events. Underproduction, on the other hand, can lead to a brown-out: a drop in voltage in the system which causes demanding systems and appliances to malfunction. Underproduction can also lead grid operators to implement rolling blackouts, where full service is only supplied to limited areas on a rolling basis. Thanks to accurate demand forecasts, and the reserved buffers of supply they inform, these underproduction outcomes are rare in nations like Australia.

There are myriad actors in the electricity market that can benefit from accurate forecasts. Electricity retailers, traders and asset managers with stakes in the energy market, and even consumers themselves can make use of predicted demand to serve their interests.

STLF has been well researched in the context of unidirectional demand, where consumers demand, and producers meet that demand. Work in unidirectional contexts benefits from a unitary focus on consumer activity analysis, with aggregate patterns of activity across the grid providing sufficient information. Of course, as electricity must be produced and consumed effectively simultaneously, aggregation of individual behaviour across the service area serves to

smooth out the observed demand. Much successful research has been conducted into modelling this smoothed demand as a single univariate time series, propagating past demand patterns into the near future. While this framing of the problem has a strong theoretical foundation, the non-linearity of demand can require complex models to forecast accurately. Alternative techniques model aggregate demand behaviours in a multivariate fashion, with a set of variables which extract the useful social information from a raw point in time. For example, year-over-year trend, weekday/workday patterns, and typical daily demand profiles can be attributed directly to their appropriate component of time with these techniques. This also facilitates inclusion of external variables which can be highly predictive of specific classes of demand, like temperature and its relationship with heating and cooling appliances.

With the introduction of renewable energy technologies, and many economic and political incentives to utilise these technologies, consumers are now able to cost-effectively generate their own power via rooftop photovoltaic (PV) solar panel installations. These and other local electricity generation technologies introduce new complexity to the forecasting task. Behind-the-meter (BTM) generation of electricity cannot be measured by electricity providers, as it occurs within the consumers' domain of electricity management. This obfuscates the true demand, making necessary a new class of net load forecasting models which account for variables beyond consumer behaviour. Weather phenomena like solar irradiation, wind speed and direction, and cloud cover could play major roles in the scale of BTM generation, and therefore net demand of consumers. To complicate matters further, demand patterns may also be affected, as consumers intentionally shift their demand toward peak BTM generation times to optimise their own net cost of electricity.

In South Australia (SA), rooftop solar systems are a significant source of BTM generation, with systems installed on about 40% of homes in the region [1]. In rare cases, which will be discussed in a later chapter, this has even lead to BTM generation beyond the total demand of the region, i.e. negative net load. It is no surprise that modelling net demand in systems like this is more difficult, requiring more data from myriad sources, and advanced methodologies to match.

In New South Wales (NSW), there are 2-3 times more rooftop solar installations than in

SA. While this is quite a significant shift from years past, proportionally it has had a lesser impact on state-wide grid activity, because these installations serve a population about 4-5 times greater [2]. Further complicating comparison, the New South Wales grid comprises both the State of New South Wales and the Australian Capital Territory. It is assumed that many other differences between these regions affect their demand patterns as well, but fall beyond the scope of this work.

Chapter 2

Literature Review

2.1 Foreign Markets

A meta-analysis of recent research into the STLF task has shown a broad range of applied methods [3]. The various methods fall into three categories: artificial intelligence (AI), statistical methods, and hybrid models. While Artificial Neural Networks (ANN) are the most commonly applied method, hybrid approaches are becoming increasingly popular as a means of extracting the advantages of each component method. With some of these methods, prediction errors can be lower than traditional time-series methods, in terms of metrics like Mean Absolute Percentage Error (MAPE). The meta-analysis found that this was the most commonly reported metric, likely due to its higher comparability across grids with significant differences in demand patterns and minimum/base load. Other metrics reported include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Average Percentage Error (APE). The most common load prediction frequency in the literature is hourly, with prediction windows ranging from one day to one week, though many researchers did not report this particular detail. Most research is conducted using private data for markets in Asia, Europe, North America, and Australia, with other areas not receiving significant focus.

Xie et al. [4] apply a Long Short Term Memory Neural Network (LSTM) followed by a Multilayer Perceptron (MLP) as a two-stage forecasting method for data from Kanto, Japan.

They show that this two-stage method is useful for separating the time-series forecasting task from any residual differences due to external variables like weather. They find that the combination of methods reduces prediction error by nearly a third, compared to isolated LSTM and MLP models.

Beichter et al. [5] use simulated demand data to explore forecast performance under different levels of supply and demand aggregation. They find that a partially aggregated strategy which forecasts supply and demand separately resulted in superior forecast accuracy, compared to a single aggregated net demand forecast or combination of many dis-aggregated generator-level forecasts.

Browell and Fasiolo [6] propose a Generalised Additive Model (GAM) model with quantile regression and conditional parametric tails. They claim this methodology is more suited to the business decision of reserve allocation. Maintaining a suitable generation buffer or reserve level is necessary to prevent black- or brown-out conditions on an electric grid, thus modelling the probability of reserve exceedance is taken to be the key task. They find that this model better captures the trade-off between the cost of additional reserve and risk of exceedance, compared to other benchmark models. While this is surely true from a business perspective, it is also valuable and more common to focus on the 50% Probability of Exceedence (POE) estimate, also known as the most probable forecast. The researchers additionally find that a grid of numerical weather prediction (NWP) statistics does not add significant value to prediction of their data from Great Britain, but concede that other feature extraction methods could make gridded NWP more valuable.

2.2 Australian Energy Market

The Australian Energy Market Operator (AEMO) conduct in-house the operational forecasts necessary to manage their grids, one of those grids being the National Energy Market (NEM) [7]. This market covers the majority of the country, including New South Wales, the Australian Capital Territory, Queensland, South Australia, Victoria and Tasmania. Other states are covered by smaller grids, like the Wholesale Electricity Market in Western Australia. To forecast demand, AEMO separate the drivers of demand into two categories: structural, which includes

factors like population and economic growth, and random, which includes weather-driven and other consumer behaviours.

Hyndman and Fan [8] emphasise the importance of probabilistic predictions in the context of long-term forecasting of demand in South Australia. The authors propose semi-parametric additive modelling of demand from driver variables like temperature and calendar effects, followed by posterior distribution estimation from simulation, scenarios, and residual bootstrapping.

Following up their previous work, Fan and Hyndman [9] again apply semi-parametric additive models to estimate relationships between demand and relevant variables. The model inputs include time/calendar-based variables, historical demand data, and temperature forecasts for specific sites within the target power system. Prediction confidence intervals are estimated using a modified bootstrap method tailored to the unique seasonality patterns in electricity demand data. The proposed methodology is used to forecast half-hourly electricity demand up to seven days in advance for the Australian National Electricity Market. Its performance is assessed through out-of-sample experiments using real-world power system data, and on-site implementation by the system operator.

The approach taken by McCulloch and Ignatieva [10] provides a basic foundation for our own analysis. They chose a parsimonious Generalised Additive Mixed Model (GAMM) fit by weighting temperature difference from the comfort level, taken to be 20 degrees Celsius, according to the time of day, which consumption activities generally follow. Using this method, their annual GAMM achieved an R^2 of 0.89, and provided an interpretable set of smooth terms to demonstrate demand patterns. Unfortunately, other forecasting error metrics were not reported. Further specification details of their GAMM model can be found in later chapters of this work.

The STLF domain poses a unique challenge in comparing past research outputs. Likely due to the wide variance in data availability and contextual application, creating or adhering to standard processes poses many issues. Choosing how much data is reasonable to use for model training or testing, appropriate methods for cross-validation of model hyperparameters, and other questions which have more definitive answers in other fields, do not present as such

in STLF. These issues and others contribute to inefficient research outputs which do not provide as much value to the actual businesses and organisations which require load forecasts. Hong and Fan [11] note such difficulties, proposing misaligned incentives between novelty of academic research and usefulness to industry as one contributing factor. More concretely, the authors survey STLF methods with a novel framing: similar day methods (e.g. k-nearest neighbours techniques), variable selection techniques, hierarchical forecasting (e.g. use of smart grid micro-data), and weather station selection.

Chapter 3

Data Description

Both the demand and weather data used in this analysis were provided by TESLA Forecasting [12]. These include the demand variables described in Table 3.1, with each listed variant of demand being the average number of Megawatts measured or estimated during the preceding time period. In general, the provided demand data covered 30-minute time periods from 2016 to 2021. Additional processing was required to unify all data to the same periodicity, as data from AEMO has instead been published for 5-minute intervals since 2021. Additionally, to fill in some missing values, supplementary demand data was sourced directly from the Australian Energy Market Operator (AEMO) [13] via scripted download.

Label	Data Type	Units	Description
`datetime`	datetime	YY/MM/DD HH:mm:ss	Date and time of observation
`total_load`	numeric	Megawatts	Average total electricity demand
`net_load`	numeric	Megawatts	Average net electricity demand
`pv_est`	numeric	Megawatts	Estimated distributed PV generation

Table 3.1: Demand Data Details

The three demand variables are distinguished in more detail as follows. `net_load` refers to AEMO's Market Management System `TotalDemand` Dispatch field, an estimate of the total demand present across the grid, which excludes transmission losses and scheduled loads. `total_load` is quite similar, with the major difference being the inclusion of scheduled

loads. On the assumption that scheduled loads need not be forecasted, we tend not to make use of this particular variable in the ensuing work. `pv_est` is the total generation attributed to rooftop photovoltaic installations, estimated via satellite image analysis. To make our research more applicable to other grids, which may not have these types of estimates available, we make sparing use of this field.

For more detail about the definition, measurement and estimation of these terms, see the documentation at `https://aemo.com.au/`.

For both the SA and NSW energy markets, 545 `pv_est` observations were found to be missing from 29 unique days. Along with some seemingly randomly dispersed missing values, large windows were missing from 2 to 26 June 2022 and 19 to 21 December 2022. Unfortunately, AEMO's publication format for this field was not as easily parsed as other fields, so these missing data were not supplemented.

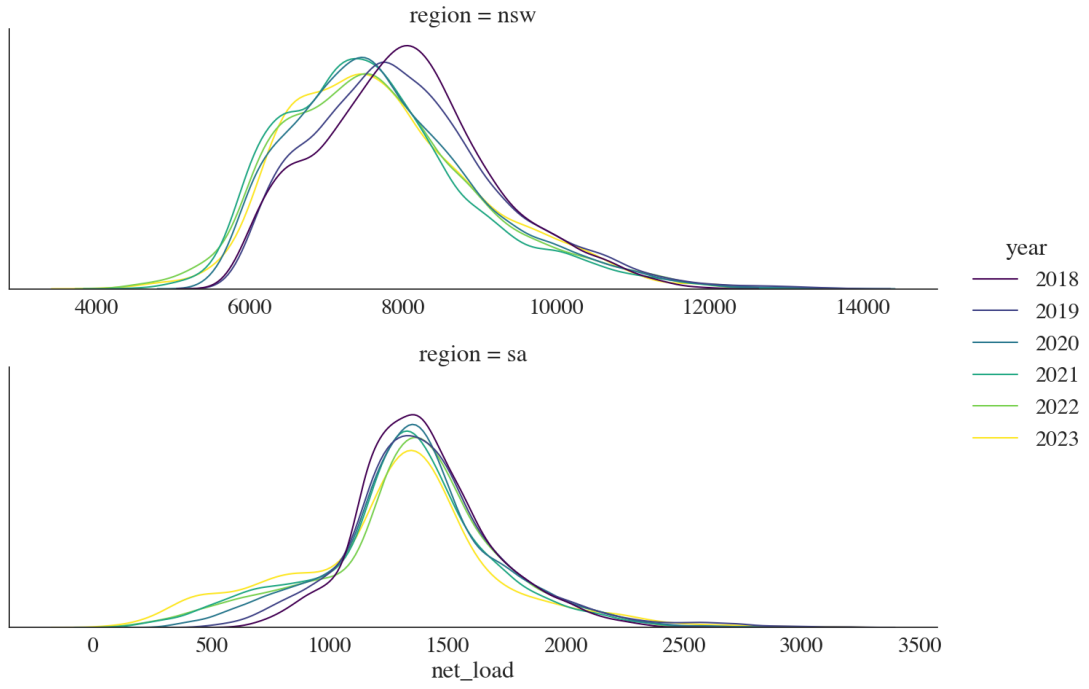

Figure 3.1: Annual Net Load Distributions by Region

The evolution of net load distributions over the available years of data is demonstrated for each region in Figure 3.1. In both regions, high load anomalies remain roughly consistent, but the majority of demand has shifted slightly down and to the left over the years, widening the distributions across the lower end. These annual shifts reflect the differential impact of changes in consumer behaviour and BTM generation in each region.

3.1 Weather Data

Hourly weather observations were provided for South Australia, New South Wales, and the Australian Capital Territory, from a single station in each of their most populous cities: Adelaide, Sydney, and Canberra, respectively. The data details are shown in Table 3.2, with the same set of variables made available from each observation site.

Histograms of each weather variable are shown in Figures 3.2, 3.4, and 3.3, for Adelaide, Canberra, and Sydney, respectively. `radkjm2` and `rainmm` have their counts on the vertical axis log-transformed for clarity; the majority of observations are zero, which precludes any insight at a normal scale. Of note in these histograms are the differences between each region's temperature distribution, especially in the lower temperature range, likely impacting differential demand for electricity to run heating appliances. Some surprising peaks also show in Canberra's humidity, cloud cover, and wind direction observations. It is assumed that these are reflective of real observations, and in the case of wind direction, attributable to a default

Label	Data Type	Units	Description
`datetime`	datetime	YY/MM/DD HH:mm:ss	Date and time of observation
`cloud8`	numeric	Oktas	Observed cloud cover
`humid`	numeric	Percent	Observed relative humidity
`radkjm2`	numeric	Watts per Square Meter	Observed solar irradiation
`rainmm`	numeric	Millimeters	Observed accumulative rainfall
`tempc`	numeric	Celsius	Observed temperature
`wdir`	numeric	Degrees from True North	Observed wind direction
`windk`	numeric	Knots	Observed wind speed

Table 3.2: Weather Data Details

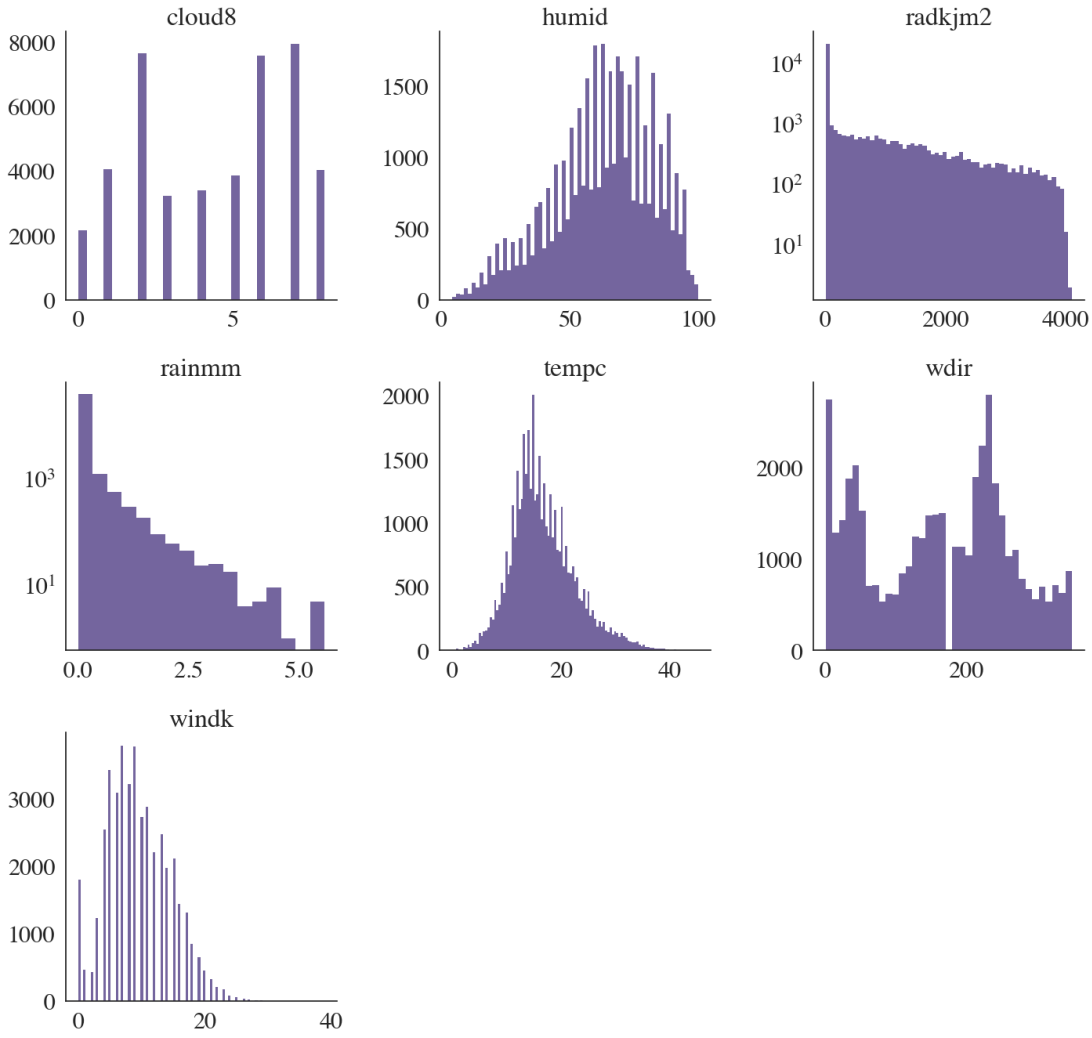

Figure 3.2: Adelaide Weather Histograms

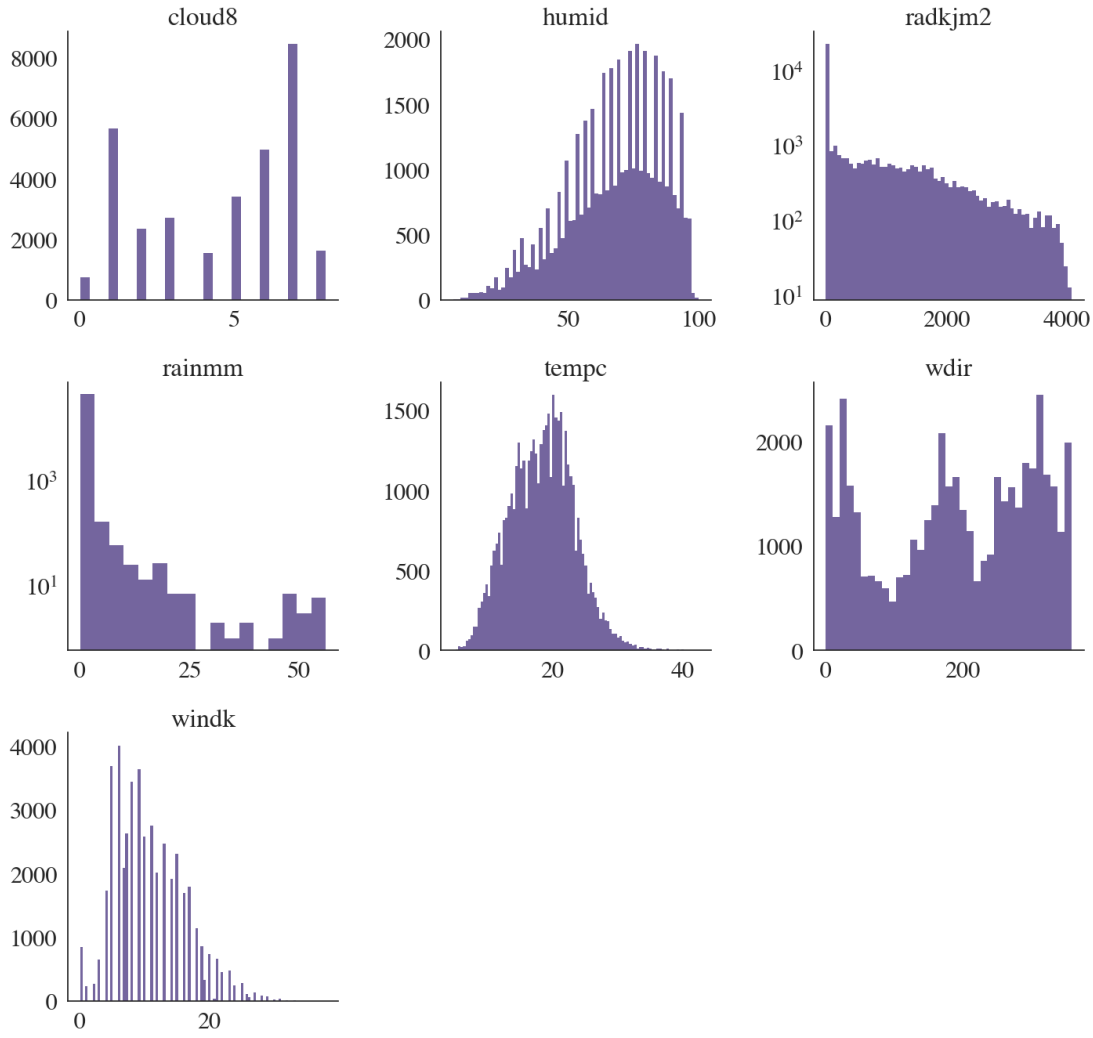

Figure 3.3: Sydney Weather Histograms

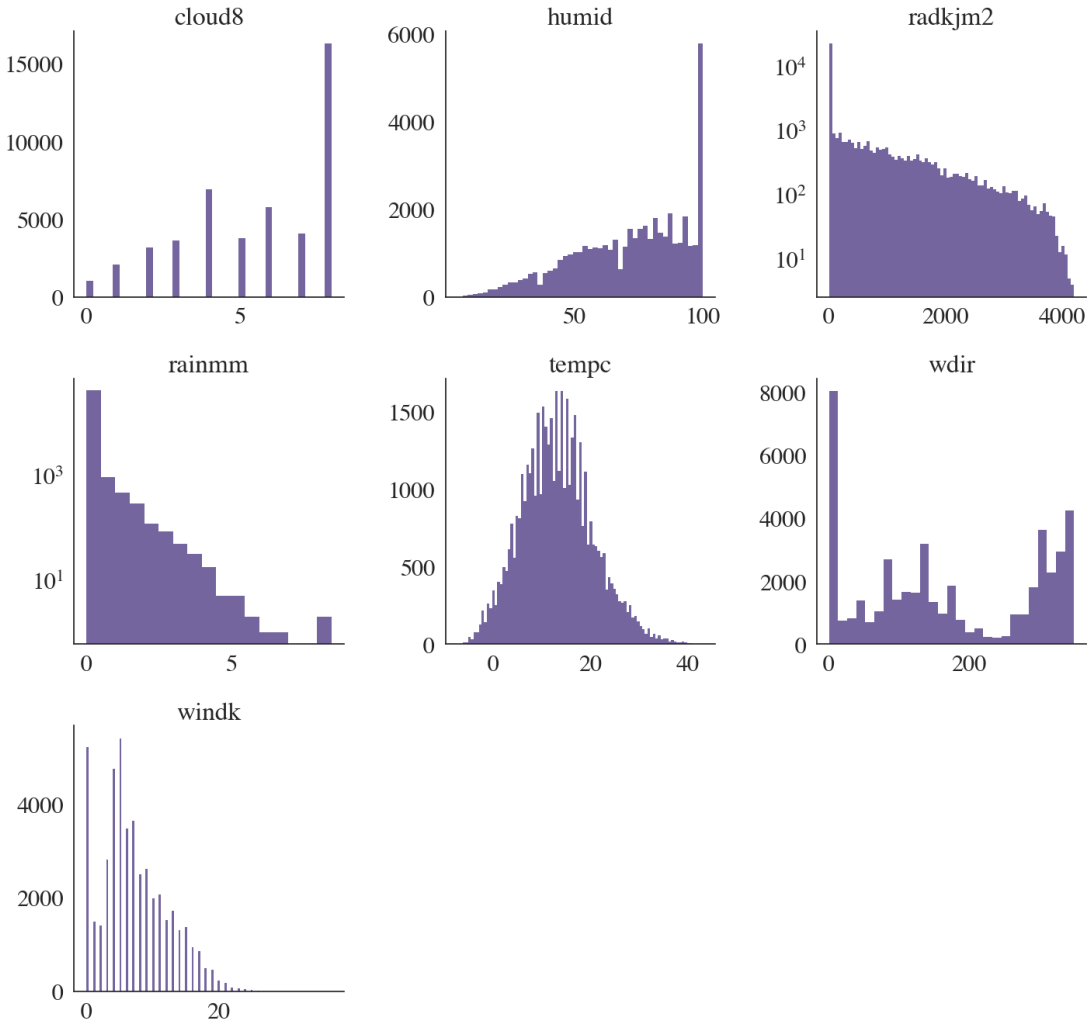

Figure 3.4: Canberra Weather Histograms

value of zero when there is no wind.

As mentioned previously, some demand observations were missing in the original dataset, but were easily retrieved from the AEMO website and slotted into their place. Some weather observations were also missing, but not so easily dealt with.

For all weather collection sites, all observations were missing from a series of 5 observations from 10am to 2pm Australian Central Standard Time (UTC+9:30), inclusive, on 9 September 2018. This suggests a malfunction of the data collection process during that window, though without a significant expected impact on our modelling, due to the narrow window of affected time. A similar situation arose with NSW weather observations in late March 2023, but as the SA data was not provided beyond early March of the same year, this did not pose an issue in our comparison.

In the weather data from Adelaide, `cloud8` and `rainmm` had 12 and 18 respective additional missing observations dispersed throughout the dataset, seemingly at random. In the weather data from Canberra, additional windows of `rainmm` were missing, though with only 30 additional observations missing, this was not to a significant degree. In the weather data from Sydney, in addition to the aforementioned September 2018 window, a significant window of weather observations were missing from most of March 2018. Interestingly, the `radkjm2` observations were not missing from this March window, but all other weather variables were. This window lies quite close to the start of the available data, so its exclusion would not have a major impact on our results. There was also a concerning proportion of `cloud8` observations missing from the Sydney data, but these were found to be mostly missing at random, and limited to observations before mid-2022.

Heatmaps of pairwise Pearson correlation coefficients between weather observations and their appropriate NEM region's net load (in the bottom row) are provided in Figure 3.5. Of note is the significant negative correlation between observed solar irradiation in Adelaide and aggregate net load for all of South Australia. This strongly indicates the scale of impact of BTM solar generation systems on net demand, confirming South Australia's unique position relative to other grids.

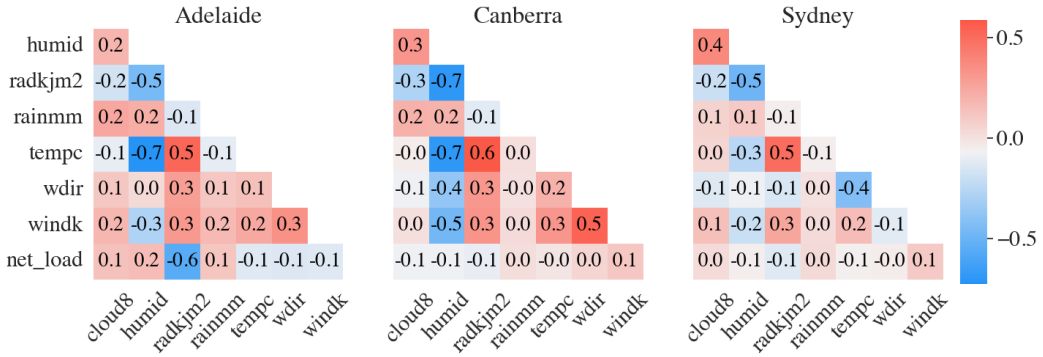

Figure 3.5: Pearson Correlations by Weather Observation Site

3.2 Pre-Processing

A unified `datetime` variable contains significant hidden information, and requires pre-processing to be suitable for multivariate modelling techniques. The year, month, day, hour, and minute (in appropriate cases) were extracted as separate variables for each model, and the original `datetime` was excluded. Additionally, numeric day-of-the-week, day-of-the-year, and week-of-the-year variables were extracted to reflect the cyclical and seasonal nature of electricity demand.

Holidays can play a major role in predictable demand anomalies. Without this additional information about societal context, the demand on weekday-holidays especially is overestimated. Therefore, a holiday dataset [14] was retrieved for inclusion in model inputs. The format of this data is quite simple: a date, name, description, and set of relevant jurisdictions (states/regions) to which each holiday applies (e.g. a state anniversary is only relevant to that particular state).

Many researchers, including those at AEMO, choose to categorise days into working days and non-working days, where non-working days include both weekends and holidays [7]. We chose to explore different strategies for encoding this information, with the methods and results

outlined in the relevant sections.

Merging the hourly weather data with half-hourly demand data required aggregating the demand into hourly averages. This resulted in a complete 1-hour interval dataset from March 2018 to March 2023, which was utilised in the following work.

Chapter 4

Data Analysis

It can be useful to demonstrate the expected pattern of cyclical time-series via profiling: averaging appropriate subsets of data across the variable of interest. The average daily net demand pattern across weekdays is one example, shown in Figure 4.1. In both regions, demand peaks during the morning and evening. In NSW, the daylight hours between these peaks shows a slight dip, with night/early-morning hours showing a more drastic dip. In SA, the magnitudes of these dips are reversed: the lowest demand on average occurs around mid-day. A clear reduction in demand can be seen through the weekend, most drastic from the morning through the early afternoon.

Recalculating this daily profile across each year available in the dataset, the year-over-year reduction in daytime net demand can be seen in Figure 4.2. This daytime reduction is accompanied by a tight similarity through the nighttime, suggesting a major change in daytime demand, but an unchanged base load. This daytime demand change is reflected at a similar absolute scale (sheer number of Megawatts) across both regions. Proportionally, however, the change in South Australia is much more significant than that in New South Wales, where base load is much greater.

Some care should be taken in the interpretation of SA's 2023 profile, as the data for that year and region spans only from January to early March.

A moving average of net load in each region is presented in Figure 4.3, with the average

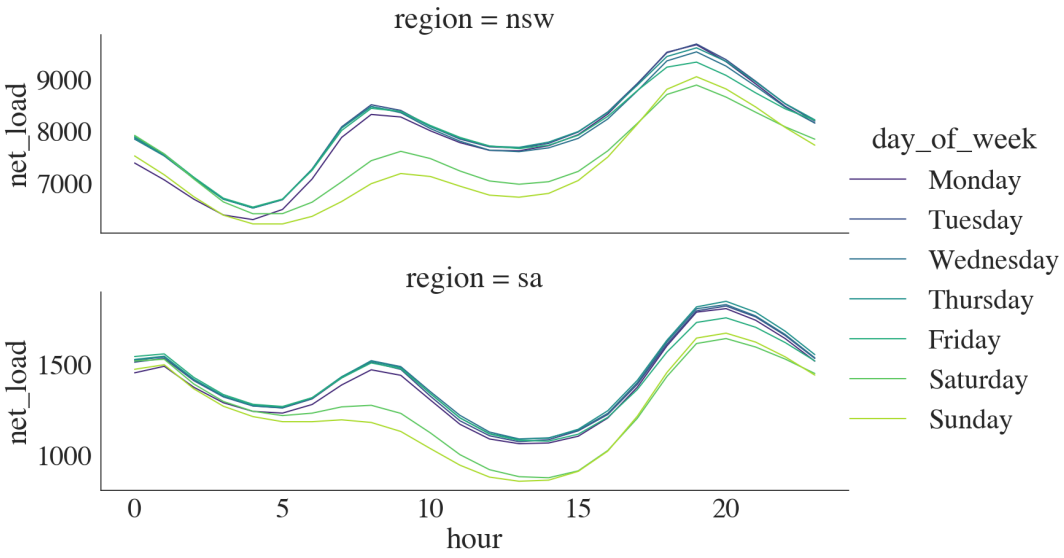

Figure 4.1: Daily Net Demand Profiles by Region and Weekday

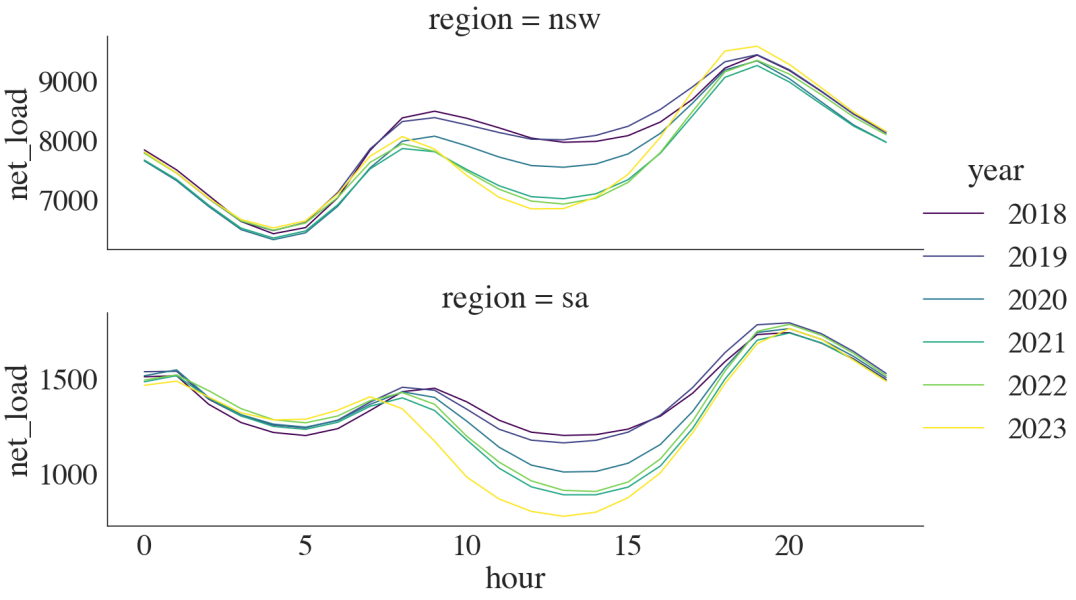

Figure 4.2: Daily Net Demand Profiles by Region and Year

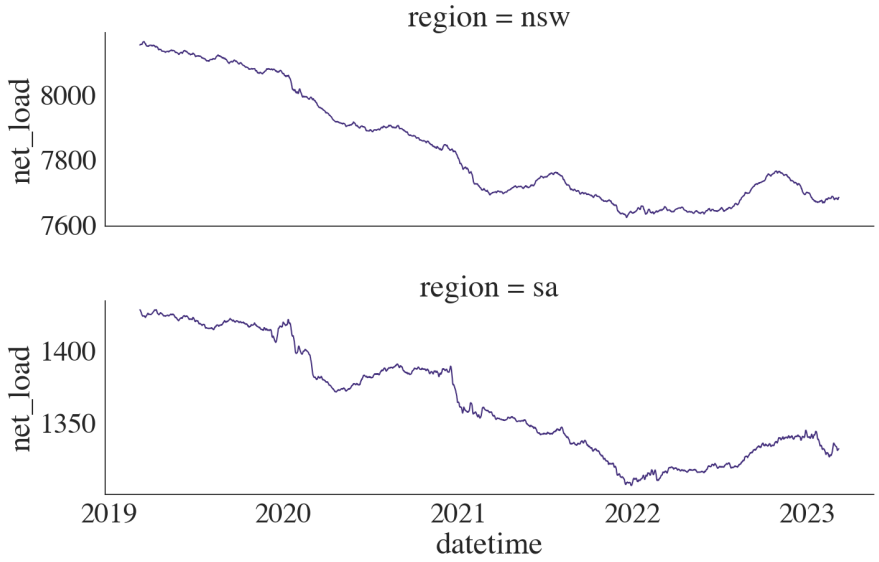

Figure 4.3: Net Load - 365 Day Moving Average by Region

covering the previous 365 days at each time point. A clear downward trend is indicated in both regions from the start of the dataset until 2022, after which there is a slight upward trend. The downward trend likely reflects increased contribution of BTM solar systems, but could also indicate consumer behaviour changes due to pandemic lockdown policies.

Figure 4.4 presents another moving average, this one covering the previous 30 days at each time point. At this level of granularity, seasonal trends can be seen more clearly. Of note are the differential changes in winter and summer demand over the years. The highest peaks of latter years show higher and smoother demand through each winter, while lower and rougher peaks reflect the more significant, but more variant contribution of BTM PV generation through the summer. In contrast, earlier years demonstrate winter and summer peaks which are more comparable in magnitude.

The non-linear relationship between temperature and net load targeted by McCulloch and Ignatieva [10] is demonstrated in Figure 4.5. Of note are the regional differences, with SA demonstrating a more consistent horizontal cluster at low temperatures that is not present in NSW. Two potential explanations for this cluster are differential usage of heating appliances

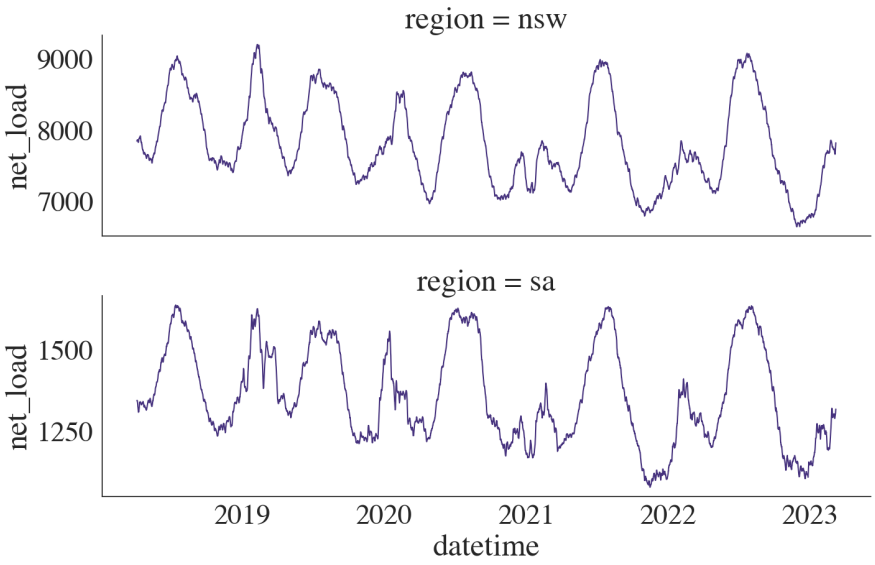

Figure 4.4: Net Load - 30 Day Moving Average by Region

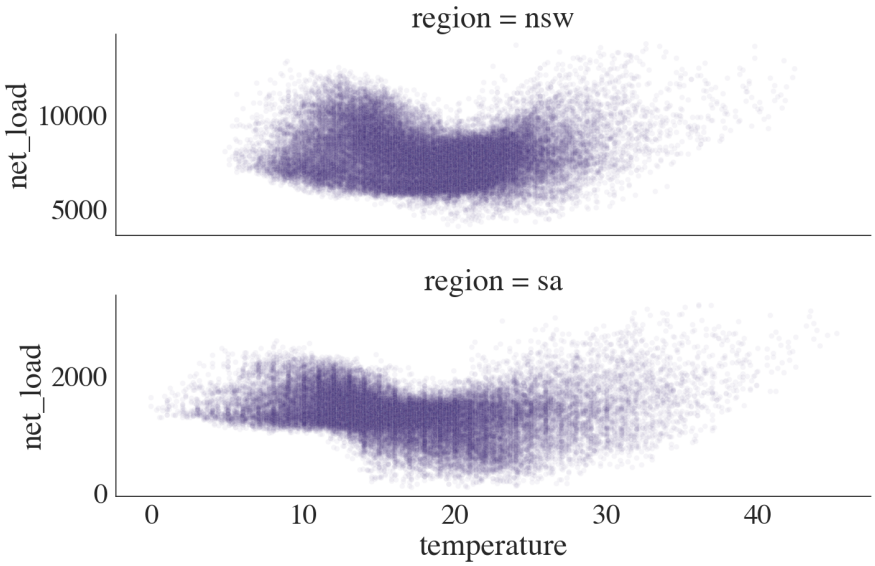

Figure 4.5: Temperature vs Net Load by Region

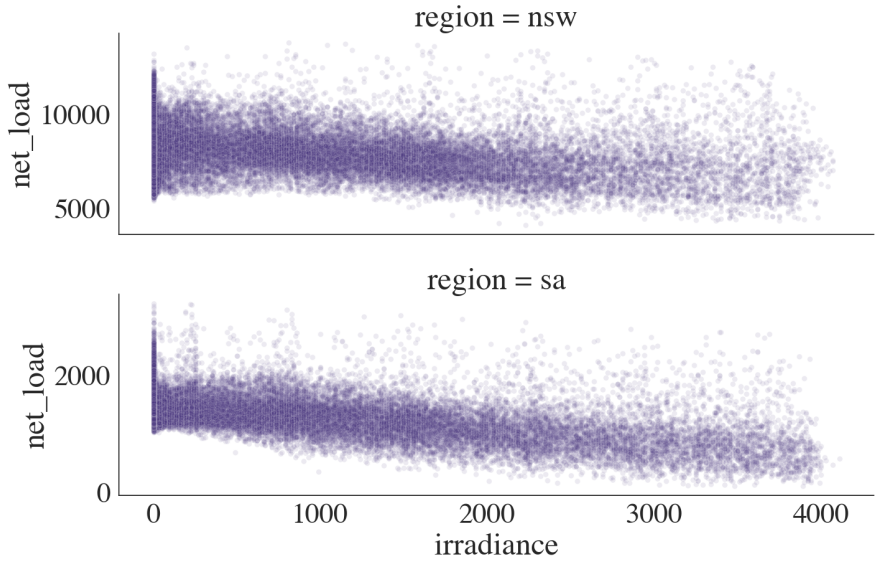

Figure 4.6: Irradiance vs Net Load by Region

or a higher proportion of days which are both sunny and cool. Sharp-eyed readers may notice vertical artifacts, especially in the SA plot; we assume these reflect temperature measurements which were rounded to the nearest whole number, for unknown reasons.

The negative linear relationship between solar irradiance and net load is shown in Figure 4.6. Interestingly, the shape of each net load distribution along the irradiance axis is fairly constant, with the majority of observations clustered toward the lower end, and a fat tail of higher net loads. Even so, there is some dispersal of net load as irradiance increases, especially in NSW, indicating less straightforward load prediction in those high-irradiance instances.

Under certain conditions in SA, the contribution of BTM PV systems has been so significant as to cause a negative total load (and near-negative net load). At these two points around midday on 21 and 27 November 2021, more electricity was added to the grid by consumers and scheduled loads than was demanded. By our analysis, this occurred under a combination of rare conditions. The base load (calculated as minimum nighttime load, typically around 2-3am) was in the 2nd percentile of nighttime loads, meaning these two days had particularly low demand even through the night, without any solar contribution. Likewise, the solar irradiation

at the time of negative load (`radkjm2` 3,500 Watts per Square Meter) was in the 4th percentile of midday (10am to 3pm) solar irradiation observations. So, these anomalies occurred by a combination of high levels of sunlight and uncharacteristically low demand. Without any known holiday effects, it's unclear what may have caused the low base load, though general public concern at the time about a new strain of pandemic disease is a strong candidate [15].

Chapter 5

Methodology

5.1 Process Considerations

As previously discussed, foundational methodological processes are not very tightly standardised in STLF. For example, McCulloch and Ignatieva [10] report an improved model fit when reducing the training set size from one year to one month of observations. It is not clear that this kind of improvement will necessarily extend to new data, or to other modelling methods. Therefore, there is some specific uncertainty around the optimal time window for model training.

The following analyses utilise both sliding-window and expanding-window training strategies, as illustrated in Figure 5.1 and 5.2, respectively. The sliding-window strategy involves fixing the training window size, for example a complete year of observations, and sliding the window along the time axis of the available data. Performance metrics averaged over each model trained in this way should reflect the comparative suitability of the chosen window size. The expanding-window strategy, in contrast, involves fixing the first observation to be included in all training sets, and progressively expanding the size of the training set to include ever more data. Aggregate performance metrics from these models should reflect the comparative suitability of incorporating additional data beyond the minimum window size.

Given these training strategies, the optimal testing strategy is slightly more obvious, but

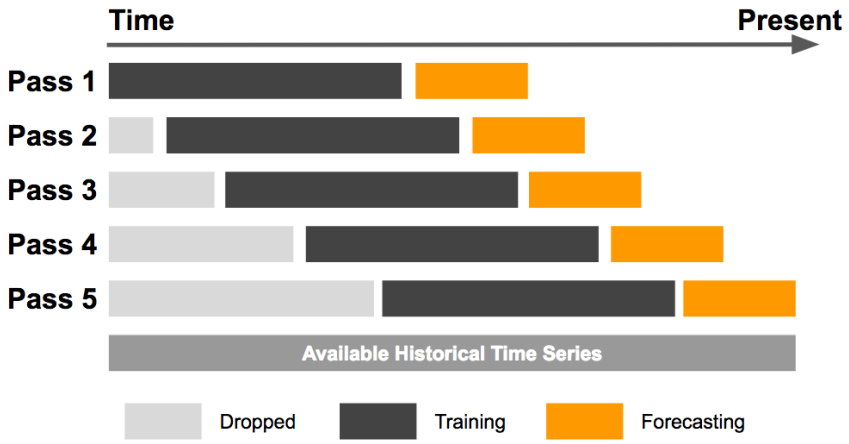

Figure 5.1: Sliding Window Strategy. Reprinted from Yang [16]

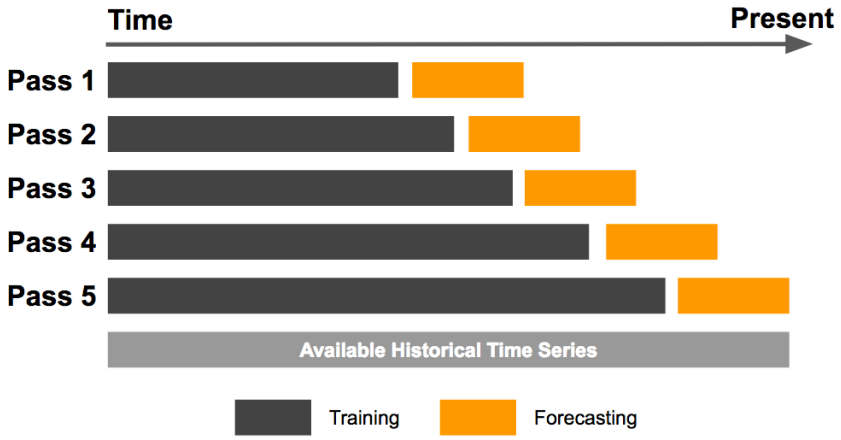

Figure 5.2: Expanding Window Strategy. Reprinted from Yang [16]

not completely so. In general, test predictions should be computed for some number of observations which directly follow the end of the training set, as they would be in actual application. Those predictions from the ensuing 24-48 hours are most relevant in the short-term forecasting domain, but, depending on the training strategy, a test window of that size makes very inefficient use of each model. To thoroughly utilise our complete 2018-2023 dataset at this level of granularity, up to 1000 models would be necessary to obtain independent predictions for each

24-48 hour period. On the other hand, forecasting based on weather observations, rather than weather forecasts, makes the use of longer test windows more tractable. While this can inflate accuracy metrics from research, relative to real operational demand forecasting, it is necessary to maintain a focused scope of research. We chose a test window of 7 days as a suitable trade-off. These and other trade-offs between accuracy, granularity, thorough use of data, and computation time are explored in later sections.

5.2 Baseline Model Replication

There are several ways to approach the STLF modelling task, the most parsimonious of which is as a univariate time series. By utilising this well-established strategy, we can assume a strong theoretical foundation while minimising the number of parameters in the model, and simplify interpretation of results. A Generalised Additive Mixed Model (GAMM) meets these requirements by fitting smooth functions to each predictor and predicting the sum of those functions' outputs for a given data point. The particular strength of this method, in comparison to other univariate time series methods, is its ability to fit mixed types of smooth curves to each covariate. This enables highly non-linear trends to be closely fit by the model.

Our GAMM baseline model adhered as closely as possible to that of McCulloch and Ignatieva [10]. Consumer activity was modelled there in a novel way: by weighting the difference between observed temperature and comfortable temperature by time of day. This is driven by the intuition that consumer demand is strongly dependent on time of day, and that temperature's impact on that demand is not linear, but one attenuated by hours of activity. For example, very low temperatures at 4am do not affect demand as much as low temperatures at 8am, when consumers are far more likely to demand electricity via heating appliances.

The implementation of this weighting strategy required normalising DST time into a $[0, 1)$ range, such that the first observation of the day accorded to 0, and the last observation just less than 1. This time-of-day encoding is input to a piecewise continuous sinusoidal function, which returns a value between 0 and 1, with 1 representing the full temperature signal, and 0 none. This weight is multiplied by the temperature difference from the comfort level of 20 degrees, resulting in a weighted temperature value which as a series can be smoothed and included in

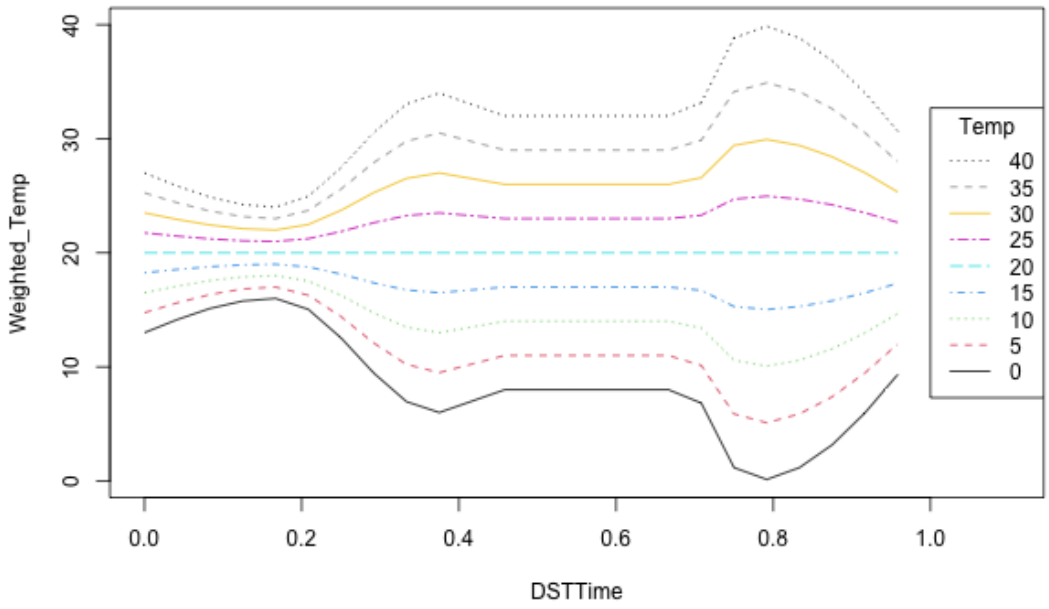

Figure 5.3: Temperature Weighting Illustration

the model. This weighting process is illustrated for several ideal constant-temperature days in Figure 5.3. As can be seen, all temperatures are down-weighted toward the comfort level in the early morning, when demand sensitivity is lowest. Dynamic attenuation follows, according to observed demand patterns, and the true unweighted temperature is returned for the evening peak, when demand for heating and cooling appliances is most sensitive.

McCulloch and Ignatieva [10]'s model was fit to weekdays in New South Wales (NSW) from March 2014 to March 2015, using Year, DST time, and time-weighted temperature. The Year and DST time data were [0-1) normalised, such that the first observation of the Year or day accorded to 0, and the last observation just less than 1. DST time was modeled as a cyclic cubic spline, a type of polynomial smoothing which ensures that the end point matches up with the start point (e.g. $y_1 = y_2$ when $x_1 = 0, x_2 = 1$). This spline type specification

should result in a better match to the true average daily cycle by ensuring continuity between days. The other terms were modeled as thin plate splines, which are particularly effective in capturing non-linear relationships between variables. In the context of spline modelling, knots are specific points in the range of a predictor variable where the curve transitions from one polynomial segment to another. The choice of the number and placement of knots affects the shape and flexibility of the resulting curve. The authors found that the optimal number of knots for DST time, DST time-weighted temperature, and Year were 12, 8, and 7, respectively.

The performance and suitability of this modelling technique may have been affected by changes in the data over the years. As a continuous dataset from 2014 to 2023 would be ideal for testing this hypothesis, further data was sourced from AEMO [13] and the weather modelling service provided by Open-Meteo [17]. This weather data is not expected to be as accurate as the weather observations in the dataset provided by TESLA Forecasting, but suits this comparative purpose well enough. This temporary dataset is only used in this section.

The training strategy here is that of an annual sliding window. Each R^2 fitness metric is computed for weekdays within a 365 day window, where the start of that window is advanced by some number of observations (roughly 5-6 months), such that 20 annual windows in total span the complete 2014-2023 temporary dataset.

The parsimonious GAMM model was fit using R Statistical Software (v4.3.1) [18] and the `mgcv` package [19] with the specification detailed above, and labelled `gamm`. In order to explore the trade-off of accuracy for parsimony in the original model, two variants of this model were also fit in tandem. A minimally modified version `gamm.rad`, which simply added direct normal irradiance as a thin plate spline term, and a maximally modified version `gamm.all`, which added all available weather variables as thin plate spline terms. All of these model specifications are provided in detail in the Appendix, and in our public code repository[1].

Comparison of how these three GAMM variants fit the data over the years should give some idea of their robustness as forecasting models. As is shown in Figure 5.4, the parsimonious GAMM has not proven very robust to changes over the years. While the maximally modified

[1] `https://github.com/chaserobertson/tesla-stlf`

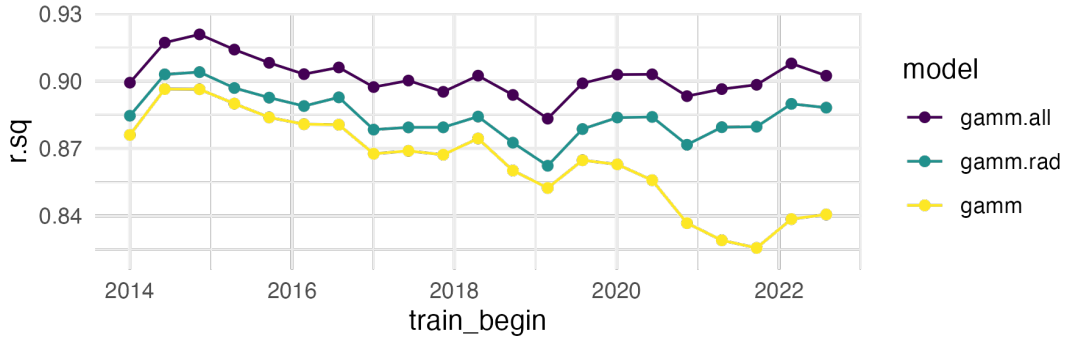

Figure 5.4: GAMM R^2, NSW 2014-2023

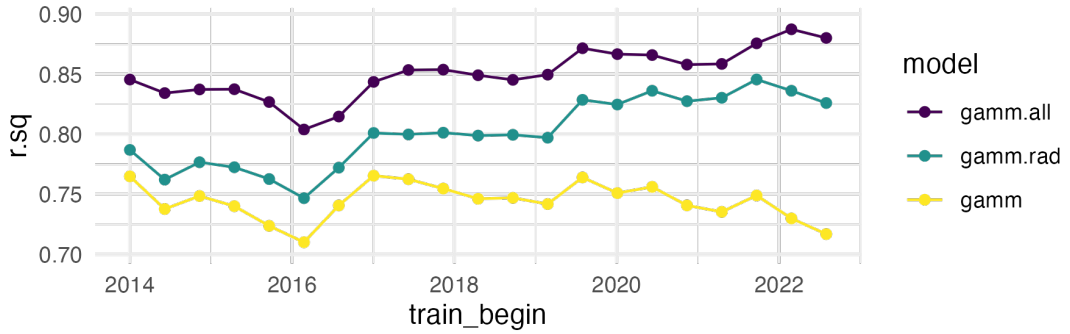

Figure 5.5: GAMM R^2, SA 2014-2023

variant achieves the closest fit across all years, the minimally modified variant keeps close pace. This close relationships suggests that the addition of direct normal irradiance maintains parsimony while providing most of the information necessary for a close fit.

When applied to SA demand data, the GAMM models in general do not perform quite as well as in NSW, as shown in Figure 5.5. It is likely that the highly customised piecewise sinusoidal weighting function used by the original authors for temperature weighting is not well suited for extension to new regions. Even so, the fitness trend of each model illustrates the information gain from irradiance and other weather variables. Again, the minimally modified

variant trends well over the years while maintaining comparable parsimony to the original unmodified GAMM.

5.3 Beyond Baseline

Though the GAMM technique does not seem to model extremely accurately in this new context, it maintains its appeal of precedence and interpretability. It is plausible that prediction performance could be improved by fitting a more flexible non-linear model, like Random Forest, to a GAMM's residuals. A Random Forest (RF) model is a Bagging ensemble method that uses independent subsets of the training data to fit a large number of decision trees, with each tree utilising only a small independent subset of the available predictor variables. Prediction is computed as an average of the predicted value from each tree. This approach is well-researched and in general robust to missing values, complex and noisy data with outliers, and overfitting. An additional advantage is independent parallel computation of each tree, and the provision of variable importance estimates based on each variable's influence in the relevant trees.

To explore the potential benefits of residual modelling in more detail, the same three GAMM variants from the previous section are fit, along with three new RF variants. One RF is a completely independent model from the others, as would traditionally be done to compare between separate methodologies. The other two RF variants, which could be considered simple applications of the boosting technique, are fit on the residuals of one of the two minimal GAMM variants. These two RFs utilise all available covariates to predict the residuals of their source GAMM, for which the available covariates are limited. Again, the model specifications are provided in detail in the Appendix and in our public code repository[2]. The data used here is not the same as that used in the previous section, which was only temporarily useful, but again the primary dataset which has been described in great detail.

To unify this attempt at modelling the residuals of the baseline GAMM model, the training regime matches that of the previous section, where 20 year-long windows were used to train each model variant. However, this case differs in performance measurement strategy. Here

[2] `https://github.com/chaserobertson/tesla-stlf`

standard prediction metrics (MAE, MAPE, and RMSE) are computed for predictions within the week of observations immediately following the training set.

The Random Forest models were fitted as described in R with the `randomForest` package [20], using all available weather variables, plus 0-6 encoded weekday identifiers. Though Random Forests tend to be robust to hyperparameter settings, variants of the size of terminal nodes were attempted, with results not varying significantly. Individual results from the South Australia data using default RF hyperparameters are presented in Figure 5.6. Of minor note are the universal spikes in error just before 2019 and 2020. These indicate particularly difficult forecasting periods during December of 2020 and 2021 (the associated test periods for those models). Given the abnormal pandemic-related social conditions around those times, unpredictable demand is not too surprising. The error spikes in latter years, in contrast, more informatively reflect the differences between models. The parsimonious GAMM does not make use of irradiance, so is unable to predict accurately through the summer. While the RF fit to its residuals does make use of irradiance, and is thus able to predict more accurately, it is not quite as accurate as the other variants. The RF model fit directly to the data performs consistently better than the others.

Model	MAE	MAPE (%)	RMSE
gamm	141.75	13.36	179.76
gamm.all	115.10	10.30	147.97
gamm.rad	113.97	10.10	149.33
rf.res.gamm	113.55	10.08	147.24
rf.res.gamm.rad	103.83	09.26	137.41
rf	94.03	08.62	129.33

Table 5.1: Residual Modelling Aggregate Metrics, SA 2018-2023

For each of the models shown in Figure 5.6, aggregate results are shown in Table 5.1. The addition of irradiance improves GAMM performance significantly, and fitting an RF model to the residuals makes some minor improvement as well. Interestingly, the simple RF fit directly to the demand data makes the most accurate predictions, across all three metrics.

Given the relative success of the Random Forest model, RF and its related ensemble methods were considered for improved modelling accuracy moving forward. For convenience of

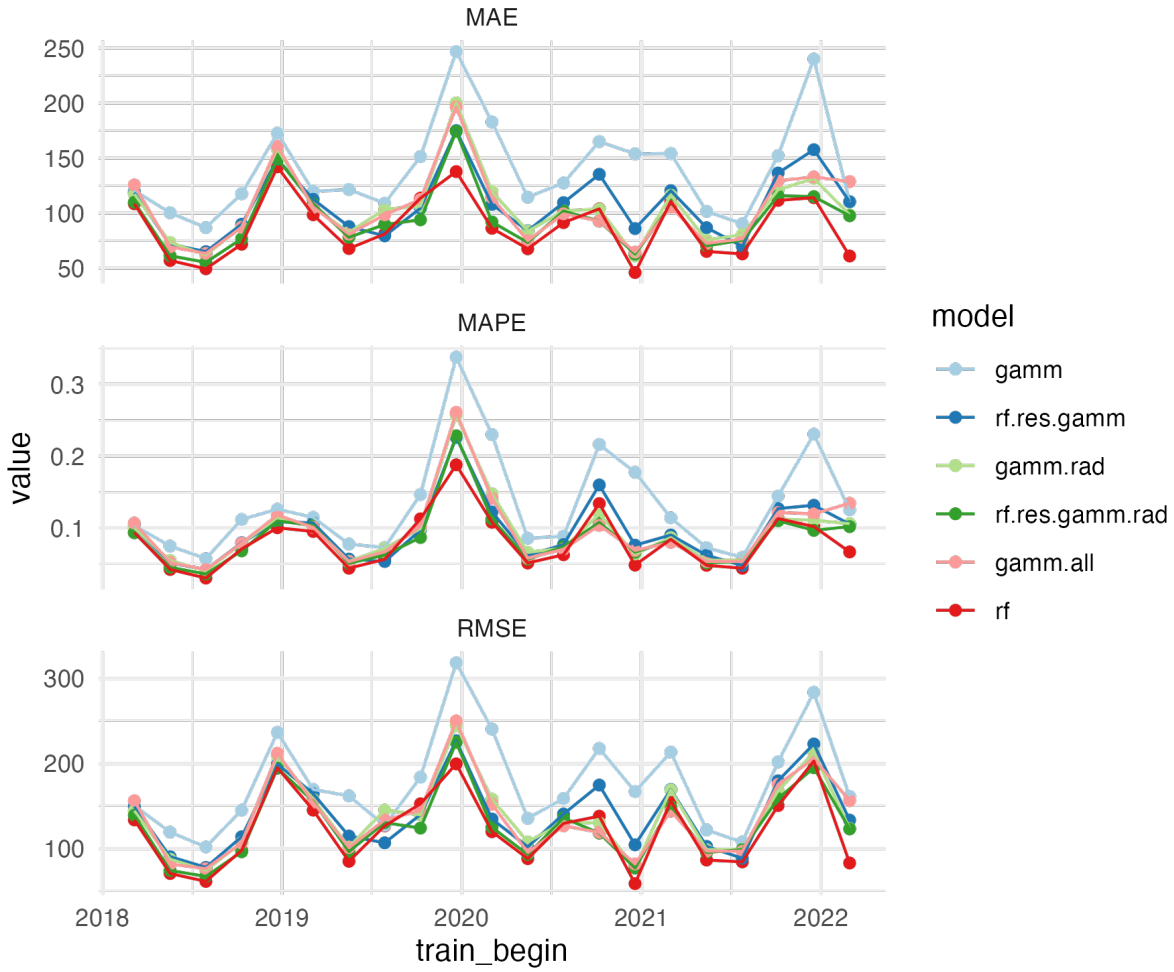

Figure 5.6: Residual Modelling Metrics, SA 2018-2023

application of a wide variety of modelling methods, the `scikit-learn` package for Python was utilised [21] for the ensuing analysis. This package provides an extensive library of model implementations and utility functions, all adhering to a standardised Application Programming Interface (API) which greatly simplifies modelling and model comparison.

One of the methods related to Random Forest is Histogram Gradient Boosting (HGB): an advanced ensemble method that refines traditional Gradient Boosting by employing histogram-based binning techniques. It transforms each individual feature space into a discrete space, then during tree construction, these discrete spaces are used to select tree splits more efficiently, reducing computational costs. It maintains the boosting architecture of sequential weak learners, while achieving competitive predictive accuracy and robustness to outliers and missing values. In our analysis of ensemble techniques available through the `scikit-learn` package, including Random Forest, traditional Gradient Boosting, and Ada-Boost, HGB struck the best balance between accuracy and training time. HGB models are used in the ensuing experiments to leverage its training efficiency and facilitate more dimensions of experimentation.

5.3.1 Holiday Encoding

To explore any interaction between holiday encoding and training window strategy, the following encoding variants were used. First, the naive encoding: no holiday information at all. Second, a simple binary encoding, where observations occurring during a holiday are tagged with 1, and all other observations 0. Third, a categorical encoding, where observations are tagged with an integer representing each unique holiday, with 0 representing no holiday. It is worth noting that around 3% of all observations occurred during a holiday, which gives some idea of the potential impact this new information could have. Finally, a binary encoding of working days was included, where 1 is the default, and 0 is assigned to observations occurring during a weekend or holiday.

Each of the holiday encoding variants was used for the SA data with both the sliding and expanding window strategies. Both window strategies started at the first observation in March 2018 with a minimum training window of 365 days and a test window of 7 days. Successive windows were 30 days apart, for a total of 49 windows yielded by each strategy. To reiterate,

Window Strategy	Holiday Encoding	MAE		MAPE (%)		RMSE	
		Overall	Holiday	Overall	Holiday	Overall	Holiday
Expanding	Binary	93.78	85.17	8.46	11.02	134.30	117.17
Expanding	Categorical	93.70	88.82	8.46	11.35	134.56	120.77
Expanding	None	93.33	146.20	8.44	17.31	133.24	190.02
Expanding	Working Day	92.25	91.82	8.27	11.50	131.92	123.37
Sliding	Binary	98.84	95.64	9.08	11.98	141.39	131.29
Sliding	Categorical	99.19	98.78	9.13	12.16	141.94	132.85
Sliding	None	97.94	146.64	8.97	17.67	140.43	192.45
Sliding	Working Day	98.20	98.06	8.99	12.71	141.30	136.27

Table 5.2: Histogram Gradient Boosting Window Strategy Metrics, SA 2018-2023

the only difference between windows yielded by the two strategies was each training window's start point, which remained fixed for the expanding window strategy, but moved ahead by 30 days for each sliding window.

For each window of each strategy, the same sub-sample of data was used to train and test a distinct HGB model for each holiday variant, yielding the aggregate metrics presented in Table 5.2. Separate metrics are reported for test observations occurring during a holiday, to clarify each strategy's effect on its goal.

While the differences in performance are not massive, they do provide some new information about each strategy and encoding. As expected, leaving out any holiday information from the models leads to the worst predictive performance on holidays. Perhaps less expected are the slightly lower figures reported by the expanding window models. This suggests that more training data does provide helpful additional information about holiday demand patterns, even if holiday demand patterns are quite different from year to year.

In addition to the previous analysis, a supplemented window training strategy was attempted. The intuition behind this particular strategy is this: there are very few holidays present throughout the year, so a standard 365-day training window would not have enough holiday observations to inform a model well. If holidays from before the general training window are included, supplementing the general window, holiday prediction may improve. The supplemented window starts with a fixed annual window, but is supplemented with data from all holidays from past years. For example, all observations which occurred during 2022 might

be included in a training set, along with each individual holiday from 2018 through 2021. We found that this strategy actually slightly degraded prediction accuracy. It is assumed that the additional information provided by the supplementation was not helpful due to the variance of holiday profiles, sensitivity to exogenous conditions, and other year-over-year changes. It could also be the case that the models did not incorporate the supplemental information in a balanced fashion, due to their discontinuous nature.

The expanding window strategy provides comparative insight, but cannot directly inform the optimal training set size for any individual prediction task. Leveraging the insight that more than a year of training data may lead to superior prediction, a similar analysis was conducted into sliding window sizes. Four sizes were attempted: semiannual, annual, biannual, and triannual, covering 182, 365, 720, and 1085 days, respectively. Across each of these sizes, the aforementioned holiday encodings were used to train distinct models and predict the ensuing 7 days as a test set. The error distributions are presented in Figure 5.7, with a row for each metric, column for each prediction category (non-holiday/holiday prediction), an adjacency group for each holiday encoding, and an individual boxen plot for each window size. As this is quite a crowded comparative analysis with many dimensions, plot artifacts deemed not strictly necessary have been removed.

While there are no statistically significant differences between any of these combinations, predictions made from semiannual training window models are clearly the worst performers. Biannual windows generally predict holidays best, though annual and triannual windows seem competitive.

Given these results, an annual sliding window strategy with binary holiday encoding is taken to be a simple, yet effective option. The ensuing analysis makes use of this methodological decision.

5.3.2 Prediction Error Analysis

Having explored modelling techniques, holiday encoding options and the general methodological space somewhat, the question remains as to why prediction performance is still relatively poor, compared to results from other research. In Figure 5.8, the standard quartiles of prediction

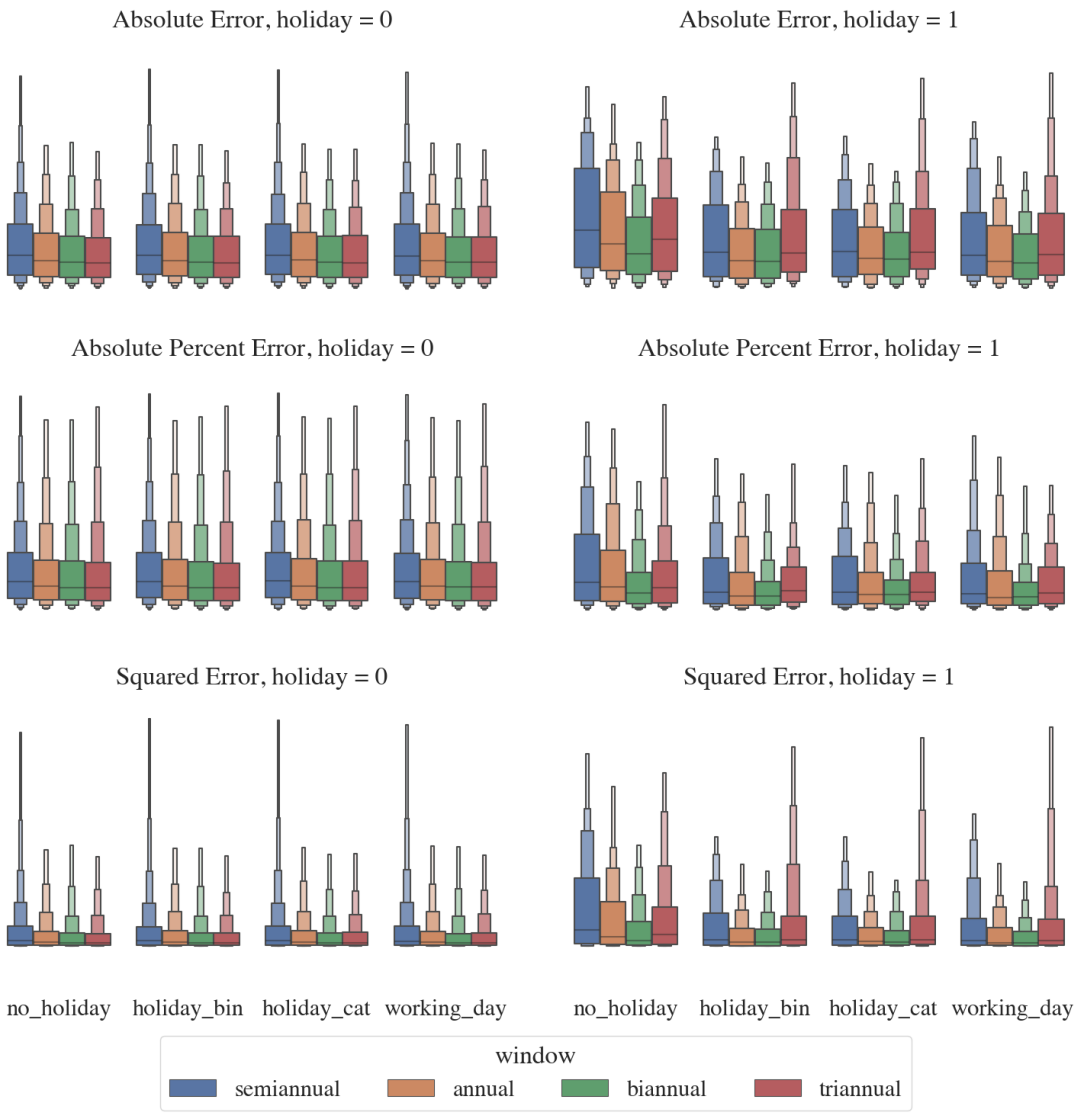

Figure 5.7: Histogram Gradient Boosting Window Size Metrics, SA 2018-2023

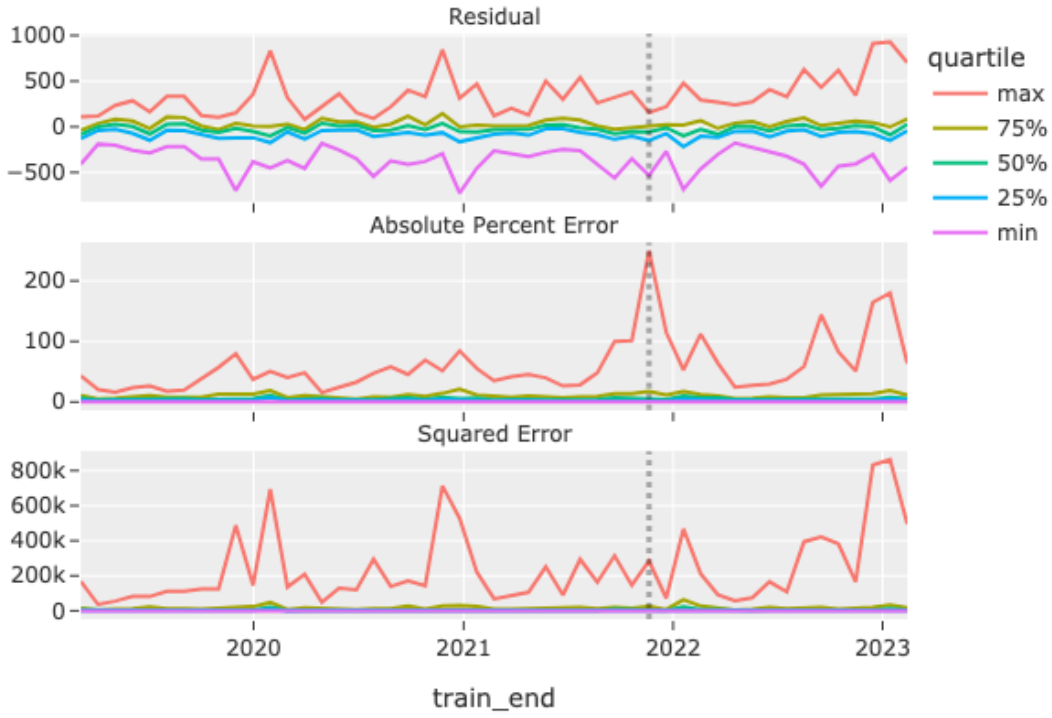

Figure 5.8: Histogram Gradient Boosting - Metric Quartiles Over Time

residuals, absolute percentage error, and squared error are presented; the resulting distributions give some hints in this regard.

A look at the residual quartile plot shows that the majority of predictions are quite near to their true value, but many extreme errors are present as well, especially through the summer months around each new year. The extreme values in this section of the plot also indicate a slight bias toward under-prediction (where residuals have a positive value), sometimes under-predicting by as much as 1000 MW.

The squared error plot quite strongly emphasises the greater difficulty of prediction through the summer. Significant but inconsistent jumps occur around each new year as one traces the maximum squared error line. While some years do present this jump around the New Year holiday period, this pattern is not universal, or at least not well captured at the granularity of

these sliding window models.

The absolute percentage error plot demonstrates quite an interesting pattern. Summertime peaks are shown, but to a relatively low extent in earlier years, with much higher peaks in recent years. The worst predictions can be so far off that they're more than double the true load! Cross-referencing between each plot, it can be seen that the greatest-magnitude residual from the highest-MAPE model (marked with the dashed vertical line) was around -500 MW: horrible prediction in relative (MAPE) terms, but not out of place in absolute (residual) terms. In addition, the predictions from that particular model are quite unremarkable in terms of squared error.

These metric disagreements highlight the difficulty inherent in measuring error in this domain. While MAPE is quite useful for comparing between different models, methodologies, and regions, it seems not to be as useful in certain applications, especially in regions like South Australia where demand can drop as low as 0 MW. When true load is near zero, even the smallest prediction error will demonstrate a ghastly, potentially infinite MAPE, as true load is in the denominator of that percentage error calculation.

To facilitate deeper insight into error patterns, the test days with highest daily sum of prediction squared error are presented in Figure 5.9, with some additional variables. True load, HGB prediction, AEMO's PV generation estimate, and observed solar irradiance are shown, labelled as `net_load`, `predicted`, `pv_est`, and `radkjm2`, respectively. The days are presented in decreasing order, meaning the greatest daily sum of squared errors was on 31 January 2020 (the top left plot), so this was the day where our methodology demonstrated its worst absolute predictive accuracy. It is not surprising that the observed irradiance appears rather jagged on this particular day, in contrast to more ideally sunny days, like 21 January 2023 at the bottom-center. What is surprising is that even with this low irradiance, the HGB model prediction is lower than the true demand, suggesting a lack of weight given to irradiance in the model. If irradiance were properly weighted, low irradiance should lead to higher demand prediction. This defect is fairly consistent across the other pictured days as well; low-irradiance days tend to be under-predicted, and high-irradiance days tend to be over-predicted.

Figure 5.10 provides more definitive evidence of the HGB models' lack of weight allocated

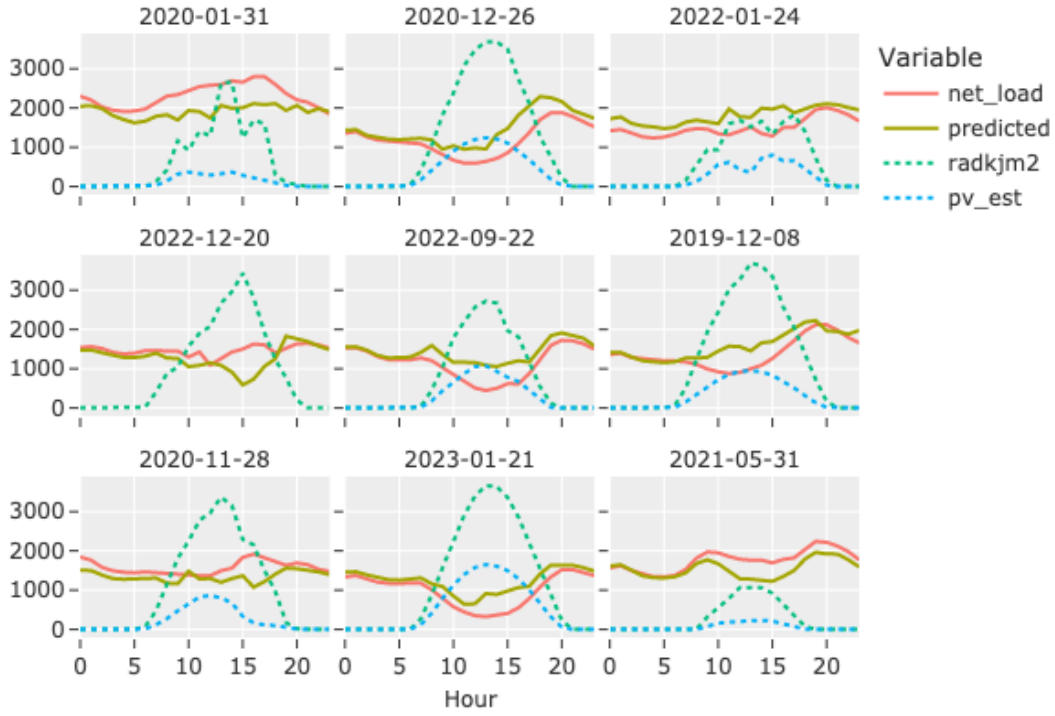

Figure 5.9: Histogram Gradient Boosting - Worst Prediction Days

to solar irradiance. This plot shows a point for each observation's irradiance value against its prediction residual, with a fitted trend line in red. The non-zero trend indicates a poor fit; as irradiance increases, the HGB models predictably underestimate the impact, i.e. overestimate the electricity demand.

5.3.3 Model Tuning

While fairly robust by default, a non-linear ensemble model like HGB can be somewhat sensitive to hyperparameter settings, so a few options were explored as outlined in Table 5.3. All performance metrics, including R^2, are from same-subsamples test sets. In addition, the `pv_est` covariate was included in each model of this exploration, to more closely simulate an applied forecasting situation in South Australia.

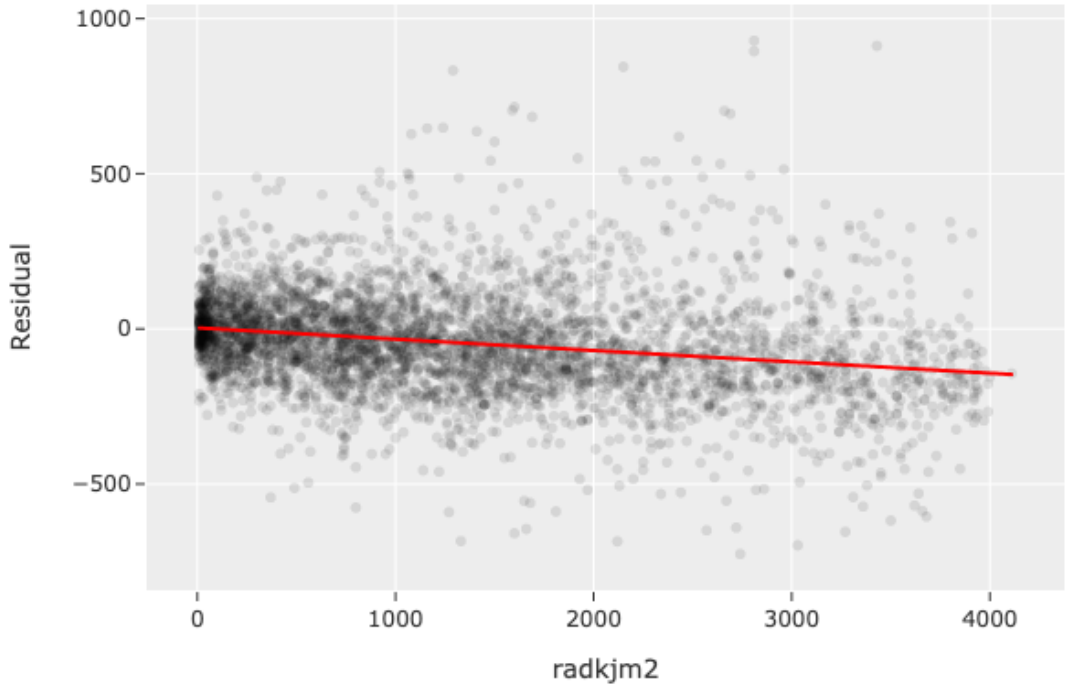

Figure 5.10: Histogram Gradient Boosting - Irradiance Vs Residuals

Various hyperparameters are available for tuning, but we limited our search to the maximum number of iterations, maximum number of leaf nodes, learning rate, and $L2$ regularisation parameters. The maximum number of iterations specifies the maximum number of trees to be fit, one tree per iteration, with a default of 100. The maximum number of leaves, defaulting to 31, restricts the number of terminal nodes in each individual tree. Learning rate, also known as shrinkage, and with a default of 0.1, is used as a multiplicative factor for the values in leaf nodes. Finally, $L2$ regularization, also known as Ridge penalty, penalises model complexity via a sum of squares measure. More details about these and other available settings can be found in the scikit-learn documentation [21].

Our search of the hyperparameter space was randomised, such that ten independent sets of hyperparameter settings were selected at random from the search space. We defined the search

space as follows: for the maximum number of iterations, either 100, 500, 1000, or 2000 was used. For the maximum number of leaf nodes, either 2, 5, 10, 20, 50, or 100 was used. For both learning rate and $L2$ regularisation parameters, a log-normal distribution with mean of 0.01 and variance of 1 was used.

These results indicate that while the default HGB model is sufficiently accurate, some increase in complexity via the number of trees or number of leaves can slightly improve performance.

Iterations	Leaves	Rate	$L2$ Reg.	MAE	MAPE	RMSE	R^2
2000	10	0.03	0.07	64.73	5.77	83.84	0.91
500	10	0.07	0.12	65.34	5.81	84.11	0.91
2000	20	0.27	0.13	68.08	6.05	90.14	0.9
100	5	0.47	0.06	74.89	6.48	96.79	0.88
2000	100	0.4	0.91	72.17	6.56	94.67	0.89
100	50	0.61	0.08	80.03	7.09	106.63	0.86
2000	10	0.78	0.02	85.23	7.6	111.31	0.85
2000	2	0.38	0.06	101.96	8.47	126.88	0.79
500	2	0.05	0.2	116.55	10.37	144.35	0.74
100	5	0.02	0.36	155.16	15.36	197.43	0.56

Table 5.3: Histogram Gradient Boosting Hyperparameter Search, SA 2018-2023

Chapter 6

Discussion

Behind-the-meter electricity generation has begun to impact grids around the world with the growth of photovoltaic and other systems. In New South Wales, this impact has been significant, but still represents a relatively small proportion of the overall electricity demand. In South Australia, BTM generation accounts for a large proportion of overall generation, which in combination with the smaller population and resulting variant demand, makes for more difficult forecasting. While much research has been conducted into the STLF task, less has been done in contexts where BTM PV is so prominent.

In our analysis of the SA data, we confirm a significant correlation between demand and irradiance, and more significant year-over-year changes in average demand profiles. Even so, we show with moving averages that both NSW and SA demonstrate similar downward trends, with increasing amplitude of seasonality.

Our application of the parsimonious GAMM published by McCulloch and Ignatieva [10] yielded mixed results. In its original form, with only time-based variables and time-weighted temperature as terms, its fitness to new data as measured by R^2 has lowered or remained low, in NSW and SA respectively. However, we found significant benefit from the addition of smoothed irradiance, and to a lesser extent the addition of other weather variables, e.g. rainfall, humidity, etc. In order to estimate these changes in performance since the original publication in a contiguous fashion, we utilised additional weather and demand data from Open-Meteo

[17] and AEMO [13].

Returning to our main dataset, we repeated our GAMM variant comparison, with the addition of Random Forest models trained either to the residuals of a GAMM variant, or to the data directly. By this analysis, we found that the summers of 2020 and 2021 were unpredictable for all models, but the summers of 2022 and 2023 were far more accurately predicted by the models which utilised irradiance information. In addition, we found that modelling the residuals of each GAMM with a Random Forest did not add significant benefit, but that fitting an RF model directly to the data yielded the best results. Based on this and a quick exploration of other ensemble methods, we chose Histogram Gradient Boosting as a suitable model for the rest of our work.

Some exploration into the interaction of data windowing strategies and holiday encoding options was also conducted. By cross-validated grid search of each strategy/encoding combination, we found that the addition of holiday information in any form was useful, but only in terms of holiday prediction itself. Overall prediction accuracy remained unaffected, due to the low proportion of holidays throughout any given year. Based on this analysis, we chose to move forward with the most simple but still effective combination: an annual sliding window strategy and binary holiday encoding.

Digging deeper into the prediction errors from the HGB models, we found significant dependence on the specific time period being predicted. This was especially pronounced through more recent summers, though absolute and relative error metrics tend to disagree on which predictions are most erroneous. While not unexpected, this variance over time does emphasise the necessity of proper methodologies when working in the STLF domain. Robust results from methods-based research require very careful consideration of the intended scope.

By analysing the 9 days least accurately predicted (in terms of squared error) by our HGB methodology, we found that irradiance seemed not to be allocated sufficient importance in prediction. Under-prediction on less sunny days and over-prediction on perfectly sunny days were fairly consistent among those days. This pattern of under-weighting was also shown in the relationship between irradiance and prediction residuals.

In our exploration of the HGB hyperparameter space, we found minor improvement by

increasing the flexibility of each model. Increasing the number of iterations and lowering the learning rate seemed to add the most benefit. Our final methodology reported cross-validated accuracies of 65 MAE, 6% MAPE, and 84 RMSE.

In conclusion, we are confident that TESLA Forecasting and other researchers will find our thorough analysis of the Australian demand and weather data helpful. We have shown that GAMM models, though appealing in their statistical foundation and parsimony, are not the most accurate predictors of demand. While ensemble methods like HGB and RF have their weaknesses, they are shown to predict more accurately given the same information. Our treatment of process considerations and inspection of prediction residuals make clear the problems which remain for STLF forecasting in solar-rich regions like South Australia.

Chapter 7

Future Work

As Hong and Fan [11] touch on in their meta-analysis, the impact of our work will likely be attenuated by a latent dichotomy in the STLF field. Due to the significant differences between grid contexts and regional demand patterns, valuable work can be conducted with a specific regional focus and limited scope. This work, best executed in tandem with grid operators or other stakeholders in the field, enables customised modelling which can be implemented and translated into live systems which provide applied value. On the other hand, more general academic work focusing on methodologies and modelling methods can provide a different class of value: foundational, if not immediately executable in a specific domain. Straddling these avenues to impact, as much past research has done, limits both the potential for applied impact and foundational information. Future research can potentially be made more valuable by focusing on one of these two avenues of impact: applied and partnership-based; or academic and foundational.

Future research in a distributed BTM generation context could potentially benefit greatly from additional exogenous information. Adding the most informative weather observations. e.g. irradiance, from several different sites across the grid region may improve forecasting accuracy significantly. With greater observational coverage of the geographical area from which solar power is harvested, forecasting models could be better able to predict the effects on net demand.

Explicit physical modelling of solar irradiation, e.g. by a combination of azimuth and number of panel installations in the region, is another potential avenue for impact. By doing so, true demand and BTM generation could plausibly be separated, simplifying the forecasting task.

More complex hybrid modelling techniques could also facilitate improved forecasting in future work. For researchers willing to pay the necessary computational costs, sequence-to-sequence or attention-based neural networks are promising avenues to explore in solar-rich grid contexts.

References

[1] D. for Energy and Mining, "Our electricity supply and market," 2023. [Online]. Available: https://www.energymining.sa.gov.au/consumers/energy-grid-and-supply/our-electricity-supply-and-market 6

[2] C. E. Regulator, "Postcode data for small-scale installations," 2023. [Online]. Available: https://www.cleanenergyregulator.gov.au/RET/Forms-and-resources/Postcode-data-for-small-scale-installations 7

[3] A. B. Nassif, B. Soudan, M. Azzeh, I. B. Attilli, and O. AlMulla, "Artificial intelligence and statistical techniques in short-term load forecasting: A review," *CoRR*, vol. abs/2201.00437, 2022. [Online]. Available: https://arxiv.org/abs/2201.00437 9

[4] Y. Xie, Y. Ueda, and M. Sugiyama, "A two-stage short-term load forecasting method using long short-term memory and multilayer perceptron," *Energies*, vol. 14, no. 18, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/18/5873 9

[5] M. Beichter, K. Phipps, M. Frysztacki, and et al., "Net load forecasting using different aggregation levels," *Energy Informatics*, vol. 5, no. Suppl 1, p. 19, 2022. [Online]. Available: https://doi.org/10.1186/s42162-022-00213-8 10

[6] J. Browell and M. Fasiolo, "Probabilistic forecasting of regional net-load with conditional extremes and gridded NWP," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5011–5019, nov 2021. [Online]. Available: https://doi.org/10.1109/TSG.2021.3107159 10

[7] Australian Energy Market Operator, "Electricity demand forecasting methodology information paper," Australian Energy Market Operator (AEMO), 2019. [Online]. Available: https://www.aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/inputs-assumptions-methodologies/2020/2020-electricity-demand-forecasting-methodology-information-paper.pdf 10, 20

[8] R. J. Hyndman and S. Fan, "Density forecasting for long-term peak electricity demand," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 1142–1153, 2010. 11

[9] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 134–141, 2012. 11

[10] J. McCulloch and K. Ignatieva, "Intra-day electricity demand and temperature," *The Energy Journal*, vol. 41, no. 3, 2020. 11, 25, 29, 31, 32, 47

[11] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207015001508 12, 51

[12] T. Forecasting, 2023. [Online]. Available: https://www.teslaforecast.com/ 13

[13] Australian Energy Market Operator, "Aggregated price and demand data," 2023. [Online]. Available: https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data 13, 33, 48

[14] A. G. D. of the Prime Minister and Cabinet, "Australian public holidays dates machine readable dataset," 2023. [Online]. Available: https://data.gov.au/data/dataset/australian-holidays-machine-readable-dataset 20

[15] A. Government, "November 2021 news archive," 2023. [Online]. Available: https://www.australia.gov.au/news-and-updates/november-2021-news-archive 28

[16] R. Yang, "Omphalos, uber's parallel and language-extensible time series backtesting tool," 2023. [Online]. Available: https://www.uber.com/en-NZ/blog/omphalos/ 30

[17] P. Zippenfenig, "Open-meteo.com weather api," 2023. [Online]. Available: https://open-meteo.com/ 33, 48

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/ 33

[19] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017. 33

[20] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/ 36

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 38, 45

Appendix A

GAMM Specifications

```
gamlwmod <- Demand ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp")

rad.form <- Demand ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp") +
    s(direct_normal_irradiance..W.m.., bs="tp")

all.form <- Demand ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp") +
    s(direct_normal_irradiance..W.m.., bs="tp") +
    s(diffuse_radiation..W.m.., bs="tp") +
    s(shortwave_radiation..W.m.., bs="tp") +
    s(windspeed_10m..km.h., bs="tp") +
    s(winddirection_10m...., bs="tp") +
    s(rain..mm., bs="tp") +
    s(cloudcover...., bs="tp") +
    s(relativehumidity_2m...., bs="tp")

fit.gamm <- function(start, form) {
    train <- subset(fitdata[start:(start + 365*24),], wday %in% 1:5)
    wtdyear <- mgcv::gamm(form,  data = train)
    summary(wtdyear$gam)$r.sq
}
```

Appendix B

Models

```
gamlwmod <- net_load ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp")

rad.form <- net_load ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp") +
    s(radkjm2)

all.form <- net_load ~ s(DSTTime, bs = "cc") +
    s(WtdTemp, bs = "tp") +
    s(Year, bs = "tp") +
    s(radkjm2) +
    cloud8 + # not suitable for smoothing
    s(windk) +
    s(wdir) +
    s(humid) +
    s(rainmm)

gam.forms <- c(gamm=gamlwmod, gamm.rad=rad.form, gamm.all=all.form)

rf.form <- net_load ~ DSTTime + WtdTemp + Year + radkjm2 + tempc +
    humid + cloud8 + rainmm + windk + wdir + wday

res.form <- residual ~ radkjm2 + tempc + humid + cloud8 + rainmm +
    windk + wdir + wday
```

```
# Sliding window methodology for one annual window with 1-week testing
fit <- function(start) {
    end <- (start + 365*24)
    train <- subset(fitdata[start:end,], wday %in% 1:5)
    test <- subset(fitdata[(end+1):(end+24*7+1),], wday %in% 1:5)
    true <- test$net_load

    # Independent RF Variant
    rf <- randomForest::randomForest(rf.form, data = train)
    pred <- predict(rf, test)
    results <- ...

    # 3x GAMM Variants
    for (name in names(gam.forms)) {
        wtdyear <- mgcv::gamm(gam.forms[[name]],  data = train)
        pred <- predict(wtdyear$gam, test)
        results <- ...

        # Also fit Residual RF Variant if applicable
        if (name != "gamm.all") {
            train$residual <- train$net_load - predict(wtdyear$gam, train)
            res.rf <- randomForest::randomForest(res.form, data = train)

            test$residual <- true - pred
            res.pred <- pred + predict(res.rf, test)

            results <- ...
        }
    }
    return(results)
}
```