Evaluating the benefit of grid-based weather information in energy forecasting

Bachelor's Thesis of

Marcel Herm

at the Department of Informatics
Institute for Automation and Applied Informatics (IAI)

Reviewer:	Prof. Dr. Veit Hagenmeyer
Second reviewer:	Prof. Dr. Achim Streit
Advisor:	Nicole Ludwig, M.Sc
Second advisor:	Marian Turowski, M.Sc

2019/07/01 – 2019/10/31

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 2019/10/31

. . . .

(Marcel Herm)

Abstract

This is an abstract. It is fine because it is small and nice.

Acknowledgements

Firstly, I would like to express my sincere gratitude to Nicole Ludwig[1], who elaborated the subject of this thesis, and Marian Turowski[2] who assisted her supervising this thesis. Their professional and incredibly considerate supervision helped me throughout my whole thesis.

To Prof. Dr. Veit Hagenmeyer[3], I am most grateful for giving me the chance to complete my thesis at IAI[4].

Further, I would like to thank Kaleb Phipps[5], who provided some very useful code, saving me a lot of time that would have otherwise been spent for debugging due to issues in the implementation phase.

In addition, I would like to thank Janine von Hodenberg for her enduring assistance and helpful consultancy and of course my family for their unquestioning support throughout many years.

[1] https://www.iai.kit.edu/2154_2161.php
[2] https://www.iai.kit.edu/2154_2288.php
[3] https://www.iai.kit.edu/Ansprechpersonen_1213.php
[4] https://www.iai.kit.edu/index.php
[5] https://www.iai.kit.edu/2154_2880.php

Contents

List of Figures

List of Tables

Terms And Abbreviations

ACF Autocorrelation Function. 20

AIC Akaike Information Criterion. 20

AnEn Analog Ensemble. 5–7

AR Autoregressive. ix, 4, 9, 22

ARIMA Autoregressive Integrated Moving Average. 4, 7

ARMA Autoregressive-Moving Average. vii, ix, 4, 7, 9, 19–22, 27

ARMAX Autoregressive-Moving Average with Exogenous Inputs. vii, ix, 4, 7, 9, 10, 21, 27

ARWD Autoregressive model using an average weekly profile. 5, 7

ARWDY Autoregressive model using an average weekly profile including annual seasonality. 5, 7

BIC Bayesian Information Criterion. 20

CARDS Coupled Autoregressive and Dynamical System. 4, 7

CRO Coral Reefs Optimization. 4, 7

DSM Demand Side Management. 1, 31

DWD Deutscher Wetterdienst. 4

ECMWF European Centre of Medium-Range Weather Forecasts. ix, 4, 5, 13, 15, 29

ELM Extreme Learning Machine. 4, 7

EMOS Ensemble Model Output Statistics. 5, 7

EPEX SPOT European Power Exchange. 4

GGA Grouping Genetic Algorithm. 4, 7

GHI Global Horizontal Solar Irradiance. 4

HQIC Hannan–Quinn Information Criterion. 20

HWT-ESM Holt-Winters-Taylor Exponential Smoothing Method. 5, 7

IDW Inverse Distance Weighting. 5, 7

KDE Kernel Density Estimation. 5, 7

LASSO Least Absolute Shrinkage Selection Operation. 4, 7

LR Linear Regression. 5, 7

MA Moving Average. ix, 9, 22

MAE Mean Absolute Error. 11, 27

MAPE Mean Absolute Percentage Error. 12, 21, 27

MARS Multivariate Adaptive Regression Splines. 4, 7

MLR Multiple Linear Regression. 7

MM5 Fifth-generation Mesoscale Model. 4

MOS Model Output Statistics. 5, 7

MPE Mean Percentage Error. 12, 27

NN Neural Networks. 4–7

NOAA/ESRL National Oceanic and Atmospheric Administration - Earth System Research Laboratory. 5

NUTS Nomenclature des Unités territoriales statistiques. 14

OK Ordinary Kriging. 3

PACF Partial Autocorrelation Function. 20

PCA Principal Component Analysis. 6, 7

PDF Probability Density Function. 5, 7

PE Persistence Ensemble. 5, 7

PV photovoltaic. 1

RAMS Regional Atmospheric Modelling System. 5

RF Random Forests. 4, 7

RMSE Root-Mean-Square Error. 11, 27

RNN Recurrent Neural Networks. 29

SSLR Simple Seasonal Linear Regression. 5, 7

SVM Support Vector Machines. 5, 7

SVR Support Vector Regression. 4, 7

VD Variance Deficit. 5, 7

WNN Wavelet Neural Networks. 4, 7

WRF Weather Research and Forecasting Model. 4

1. Introduction

TODO mention that the idea for this thesis originated from Nicole Ludwig[1]

TODO mention Demand Side Management (DSM) somewhere

According to Li et al. (2009), especially temperature and perceived temperature have a great impact on energy demand. Consequently, several authors combine forecasting energy time series using weather data. They mostly either focus on forecasting photovoltaic (PV) electricity generation as in Bofinger and Heilscher (2014) and Sperati et al. (2016) or on electricity generation from wind as in Davò et al. (2016) and Alessandrini et al. (2015).

It is notable, that works using station-based data often try to do some sort of geographic interpolation to be able to obtain values for every possible position. Considering this thesis, there is no such problem, as the used grid-based data already provides such distributed values. Thus, when using grid-based data, the step of interpolation can be omitted, and therefore, less effort is required. Some works also regard only forecasting for specific locations or accumulated values for bigger areas. With grid-based data, more general predictions can be made regarding the target location, as there is no binding to a certain locality.

Another interesting point is, that the works that forecast weather related time series all used grid-based data, though some of them also used station-based data to refine their forecasts. Among the papers that aimed for forecasting power or similar often only station-based data is used which leads to the assumption that less effort has been made in these fields, as it is still more complex to acquire grid-based data. However, there remains the possibility that station-based data is more suitable, even though this means a trade-off in terms of flexibility. Of course it is also possible that this has to do with the fact, that there is no grid-based power data available as this may harm privacy issues.

[1] `https://www.iai.kit.edu/2154_2161.php`

2. Related Work

This chapter gives an overview of related work in the field of energy forecasting considering grid-based data. After describing the general approach of search, there are some sources presented in ascending order of degree of relation to this thesis.

In order to find relevant literature, arXiv[1],Google Scholar[2] and BASE[3] are used.
In the process of search, the following criteria are applied to identify relevant literature:

- The title of the paper suggests that the authors work with geographic or grid-based data
- The title of the paper implies that the subject of the paper is being situated in the field of energy networks
- The title of the paper suggests that the authors aim at forecasting values
- The abstract or introduction of the paper suggests that the authors work with geographic or grid-based data
- The abstract or introduction of the paper suggests that the authors aim at forecasting or rather how forecasting is done

Subsequently, there will be two papers outlined which provide useful information for related research. After that, the papers that meet above criteria are outlined and explained regarding the used type of data, the forecast time series, applied forecasting methods, the location and the forecast horizon.

There is a high correlation between internet traffic load and electricity load, as can be read in Morley et al. (2018). This is why Kamińska-Chuchmala (2014) are considered to contain valuable information. They apply Ordinary Kriging (OK) to spatially interpolate station-based data. Due to the high similarity between internet traffic load and electricity load, it

[1]https://arxiv.org/
[2]https://scholar.google.de/
[3]https://www.base-search.net/

has potential for related subjects and also influenced this work e. g. concerning further research. Furthermore, this work underlines the benefit of grid-based data by illustrating the necessary effort of processing station-based data. Another suitable example utilizing station-based data is Fairley et al. (2017) which investigates marine electricity generation and critically discusses implications for electricity supply. This combines localization issues and the electricity network. Unfortunately, the aim here is not to forecast, but only to examine the problem.

The first group of related work utilizes grid-based, station-based or both types of data to forecast various, not necessarily electricity generation related, time series. In the first paper elaborated by Ludwig et al. (2015), the use of station-based weather data from Deutscher Wetterdienst (DWD) for electricity price forecasting in Germany is investigated. The price history is obtained from European Power Exchange (EPEX SPOT). The work compares Least Absolute Shrinkage Selection Operation (LASSO) and Random Forests (RF) in addition to Autoregressive-Moving Average (ARMA) and Autoregressive-Moving Average with Exogenous Inputs (ARMAX) models. A desirable side effect from RF is the output of the variable importance which is useful in order to filter variables by order of their relevance. As this work has a focus on short-term forecasts, the forecast time series here is the electricity price for the next day, thus having a forecast horizon of 24 hours. Another example of this group is Salcedo-Sanz et al. (2018), where grid-based weather data is used to forecast solar radiation in Australia. The evaluated methods are combinations of Coral Reefs Optimization (CRO), Extreme Learning Machine (ELM), Grouping Genetic Algorithm (GGA), Multivariate Adaptive Regression Splines (MARS), Support Vector Regression (SVR) for a forecast horizon of 24 hours. Similarly, Diagne et al. (2013) utilizes grid-based weather data for solar radiation forecasting. Different data sources are compared, specifically European Centre of Medium-Range Weather Forecasts (ECMWF), Fifth-generation Mesoscale Model (MM5) and Weather Research and Forecasting Model (WRF). The paper focuses on Autoregressive (AR) methods including ARMA, Autoregressive Integrated Moving Average (ARIMA) and Coupled Autoregressive and Dynamical System (CARDS), Neural Networks (NN) and Wavelet Neural Networks (WNN) considering short time ranges from 5 min up to 6h.

The second group of related work forecasts electricity generation and makes use of station-based data. E. g. Aguiar et al. (2016) utilize both, grid-based and station-based weather data to improve Global Horizontal Solar Irradiance (GHI) forecasts on Gran Canaria Island. GHI is similar to solar radiation that is forecast in the last paper of the previous group. In order to obtain the desired results, NN are applied. As the authors consider intra-day forecasting, the forecast horizon is limited to a range from 1 up to 6 hours in this case. Bofinger and Heilscher (2014) acquire data only from local weather stations to forecast solar power generation. The data is then refined with grid-based data from ECMWF by applying Model Output Statistics (MOS) and Inverse Distance Weighting (IDW), spatially interpolated and then simulated for Germany in order to predict a temporal range of 24-120 hours. In a work, that has been published by Haben et al. (2018), station-based weather data is applied to forecast low voltage load in the United Kingdom. They implement Kernel Density Estimation (KDE), Simple Seasonal Linear Regression (SSLR), Autoregressive model using an average weekly profile (ARWD), Autoregressive model using an average weekly profile including annual seasonality (ARWDY) and Holt-Winters-Taylor Exponential Smoothing Method (HWT-ESM) and compare them for forecast horizons of up to 4 days. Alessandrini et al. (2015) utilize non-gridded wind and power data from a wind farm in northern Sicily in Italy, with which they forecast generated wind power. Here, a novel approach, an Analog Ensemble (AnEn), is applied to the data to retain a probabilistic prediction for the next 0-132 hours. This method originally is used for meteorological ensemble forecasts.

The last group of related work forecasts electricity generation, but in contrast to the first two groups, only uses grid-based data. An application of grid-based data from ECMWF is proposed in Sperati et al. (2016) for solar power prediction in Italy. They implement a Probability Density Function (PDF) combined with NN, Variance Deficit (VD), Ensemble Model Output Statistics (EMOS) and Persistence Ensemble (PE). The time series forecast includes a range of 0-72 hours. Similar to this thesis, De Felice et al. (2015) use grid-based data from ECMWF to forecast the electricity demand, though for Italy. Linear Regression (LR) and a Support Vector Machines (SVM) are applied. Given that power prediction is a rather complex problem, the non-linear SVM performs better than a simple LR. The last, and therefore most relevant paper presented, is Davò et al. (2016) who utilize grid-based wind speed data generated by applying the Regional Atmospheric Modelling System (RAMS) with boundary conditions from ECMWF. Furthermore, they acquire grid-based data of solar radiation energy per square meter as one of the two forecast time series is the solar irradiance. The data is coming from National Oceanic and Atmospheric Administration - Earth System Research Laboratory (NOAA/ESRL) and was provided for an online competition hosted by Kaggle[4], an online community for data scientists providing competitions which are mainly based on machine learning tasks. Reference power data is

[4]https://www.kaggle.com/

obtained from Terna[5], one of the main European electricity transmission grid operators, since the other predicted time series is the wind power produced over Sicily. A Principal Component Analysis (PCA) is employed, as grid-based data is even more prone to the curse of dimensionality because of the two additional dimensions. In terms of forecasting, they apply NN and an AnEn. The forecast horizon has a range of 0 to 72 hours and the output is a prediction of both, wind power and solar radiation.

Comparing the works above to this thesis, it is notable that none of them focuses on the exact issue of evaluating the benefit of grid-based data to forecast energy time series. This is why other papers containing similar subjects are consulted as information sources in order to solve this problem.Table 2.1 provides an overview about the mentioned related works regarding the type of the used weather data, used methods, the place of origin of the data, the forecast horizon and the forecast time series.

[5] https://www.terna.it/

Table 2.1. Related works with type of used weather data, used methods, place of the data acquisition, forecast horizon and forecast time series.

paper	type of weather data	forecast time series	location	methods	forecast horizon
Ludwig et al. (2015)	station-based	energy prices	Germany	ARMA,ARMAX,LASSO,RF	24h
Salcedo-Sanz et al. (2018)	grid-based	solar radiation	Australia	ELM,CRO,MARS, MLR,SVR,GGA	24h
Diagne et al. (2013)	grid-based	solar radiation	-	ARMA,ARIMA,CARDS, NN,WNN	5 min-6h
Aguiar et al. (2016)	mixed	solar radiation	Gran Canaria Island	NN	1-6h
Bofinger and Heilscher (2014)	mixed	solar power	Germany	MOS,IDW	24-120h
Haben et al. (2018)	station-based	low voltage electricity load	United Kingdom	KDE,SSLR,ARWD, ARWDY,HWT-ESM	up to 4 days
Alessandrini et al. (2015)	station-based	wind power	Sicily	AnEn	0-132h
Sperati et al. (2016)	grid-based	solar power	Italy	PDF,NN,VD,EMOS,PE	0-72h
De Felice et al. (2015)	grid-based	electricity demand	Italy	LR,SVM	1-2 months
Davò et al. (2016)	grid-based	wind power,solar radiation	Sicily	PCA,AnEn,NN	0-72h
This thesis	grid-based	electricity load	Germany	LR,ARMA,ARMAX,PCA	1h

3. Methodology

This chapter introduces the methods that are applied in this thesis. The first group of methods comprises the used forecasting methods, the second group involves feature selection techniques and the last group covers methods that are used to evaluate forecasts.

3.1. Forecasting Methods

For time series forecasting, often used methods are e. g. ARMA models as mentioned in Hyndman and Athanasopoulos (2018). This sections will introduce the methods that are applied in this thesis to forecast the electricity load.

ARMA

The ARMA model is a combination of Autoregressive (AR) and Moving Average (MA) terms. The formal description is given by

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \rho_j \epsilon_{t-j} + \epsilon_t \ , \tag{3.1}$$

with c as a constant, ϵ_t as noise term with respect to time t, p as size of the AR part, q as size of the MA part, ϕ and ρ for the AR and MA coefficients respectively and y_t as the response variable.

ARMAX

An extension of ARMA is ARMAX, which includes an additional term for exogenous variables. This term can be used to include relations to external factors that do not depend on the endogenous data. It is formally described as

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \rho_j \epsilon_{t-j} + \sum_{k=1}^{n} \eta_k x_k + \epsilon_t \ . \tag{3.2}$$

The only difference between Equation (3.2) and Equation (3.1) is the additional term $\sum_{k=1}^{n} \eta_k x_k$ for the ARMAX for n included exogenous variables x with η as the respective coefficients.

3.2. Feature Selection Techniques

Because the used weather data is grid-based, there are two more dimensions than usual, where only one value per time step exists for a variable. This is why feature selection here is more important in order to obtain a reasonable computation time. In the following, the used methods for feature selection are presented.

Naive approach

First, naive techniques are presented, that are used to reduce the huge amount of grid-based weather data. They are reduced along the two spatial dimensions, longitude and latitude, for each step in time, respectively. These are simple functions such as the maximum or the mean. An exemplary formula for reducing the data along longitude and latitude using the mean is given as

$$x_t = \frac{1}{l \times m} \sum_{i=1}^{l} \sum_{j=1}^{m} x_{ij} \ , \tag{3.3}$$

where x_t is the calculated mean for time t, l and m are the size of the data along the axis of the longitude and latitude and x_{ij} is the value of a weather variable at the grid point with longitude i and latitude j.

Using Population Data

Another method involves population data from Eurostat[1]. It contains the population of NUTS 3 level regions. The regions are sorted by population and those with the highest population are used to filter the respective grid points that are then used as exogenous variables.

3.3. Forecast Evaluation

In order to estimate whether the used model performs well, it is important to apply suitable metrics to evaluate the results. In the following, the four used metrics are introduced, where for each metric, k is the number of forecast values, y the actual values and $\hat{y}$ the predicted values.

Root Mean Squared Error

The first metric is the Root-Mean-Square Error (RMSE), which is an often used, scale-dependent accuracy measure that calculates the root of the squared mean of the differences between the forecast and the actual values. It is described by

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}\ . \tag{3.4}$$

Mean Absolute Error

The second metric is the Mean Absolute Error (MAE), which is another scale-dependent accuracy measure that averages absolute errors. The equation for the MAE is described by

$$MAE = \frac{1}{k}\sum_{i=1}^{k}|y_i - \hat{y}_i|\ . \tag{3.5}$$

[1] `https://ec.europa.eu/eurostat/data/database`

Mean Percentage Error

The third metric is the Mean Percentage Error (MPE), which is a relative measure of the prediction accuracy. Since it is multiplied by 100 after dividing it by the size of the predictions, it is called a percentage error. The equation is

$$MPE = \frac{100}{k} \sum_{i=1}^{k} \frac{y_i - \hat{y}_i}{y_i} \ . \tag{3.6}$$

Mean Absolute Percentage Error

The fourth metric is the Mean Absolute Percentage Error (MAPE), which is similar to the MPE, but takes the absolute value of each single error instead. The equation for the MAPE is given by

$$MAPE = \frac{1}{k} \times 100 \sum_{i=1}^{k} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \ . \tag{3.7}$$

4. Evaluation

This chapter provides information about the input data acquired from different sources, the implementation, the results of the electricity forecasts, how they were obtained and an evaluation considering the performance and the quality of the output.

4.1. Data

In this section, the different types of data that are used within this thesis are presented. After that, the process of data preprocessing is briefly examined.

ECMWF

The data used in this thesis originates from ECMWF, which is a research institute that produces global numerical weather predictions and other data. It is time series based and for each timestamp there is a 2-dimensional array referred to by longitude and latitude respectively.

It must be mentioned, that, as the data that is used, has been reanalysed. This means, that the expected error is likely to be smaller than if working with real-time data.

As data parameters, there are also longitude and latitude, where the longitude is chosen to be from 5.5 to 15.5 and the latitude from 47 to 55.5. As the resolution of the used grid is at 0.25 degrees, this results in a total of 1435 grid points per timestamp. As the range of the data from ECMWF extends from 2015/1/1 to 2018/12/31, there is a total of 1461 days with each 24 timestamps due to the 1 hours frequency and thus 35064 timestamps. Considering that there is a value for each point in the grid and every timestamp, there are 50316840 values for each variable and thus 654118920 values for 13 variables.

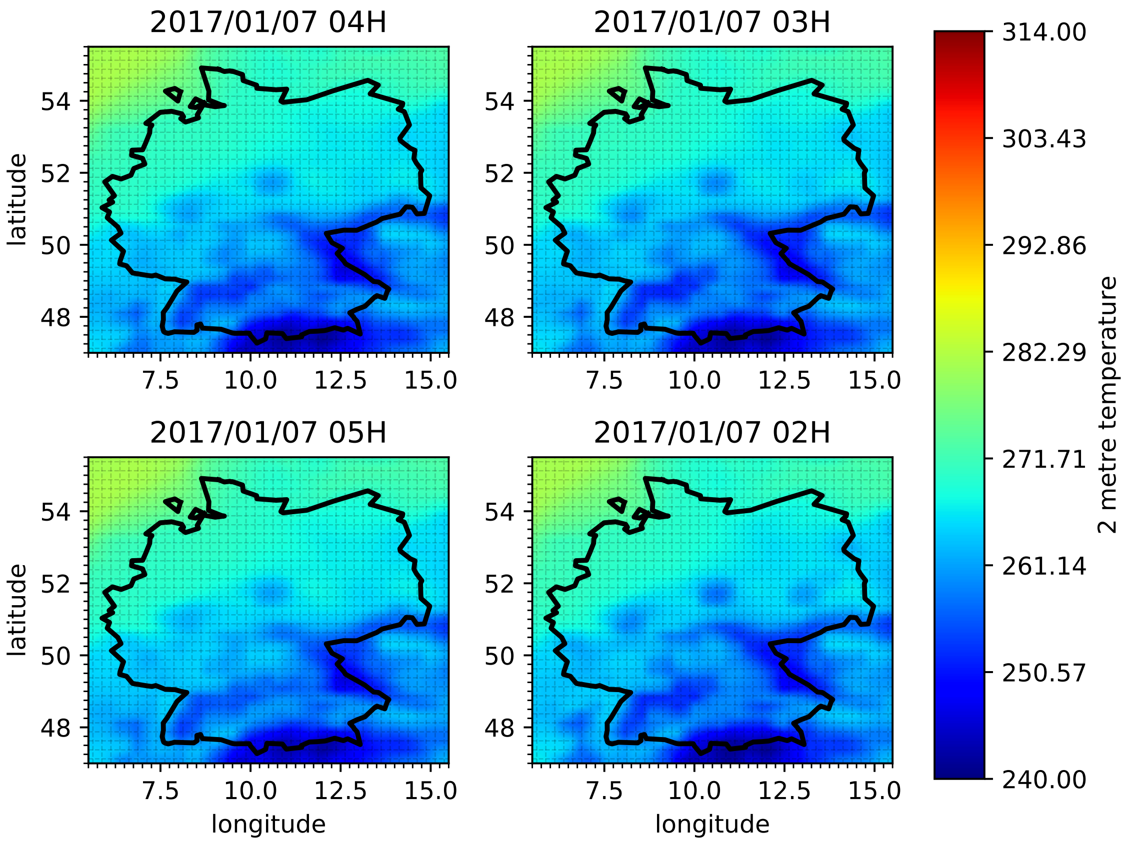

Figure 4.1. Four maps showing the four times with the highest 2 metre temperature variance in Germany, where top left is the highest, top right the second highest, bottom left the third highest and bottom right the fourth highest variance.

In order to reduce complexity, a shapefile of the NUTS dataset was used. The shapefile contains all countries in the EU. The shape of Germany was filtered from this data and each point in the dataset is checked whether it is within Germany or not. The result can be seen in the left figure in Figure 4.2. An example for a complete map without filtering is shown in Figure 4.1, where a map of the four times with the highest variance in 2 metre temperature are displayed, respectively. The filtered map first is stored in a numpy.ndarray and then applied on the data to mask unwanted data. The filtered map is visualized in the right figure in Figure 4.2. The same process was made for NUTS 3 level districts which later in this section considering population data.

[1] `https://ec.europa.eu/eurostat/de/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts`

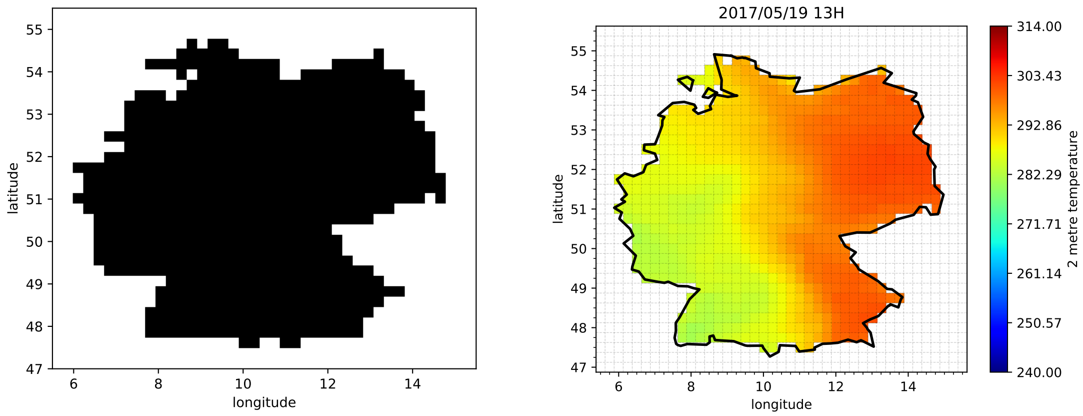

Figure 4.2. 2D boolean numpy.ndarray (left) used to filter grid squares that are within Germany. It was created by using a shapefile of Germany from Eurostat[1]and checking for each point of the grid whether it is within the shape. When applied to the weather data, only relevant data within Germany is obtained (right).

The initial dataset contains a set of variables listed in Table 4.1 where also the units, mean, min and max are shown for each variable respectively.

Table 4.1. List of exogenous weather variables used to forecast the load including min, max values from ECMWF[2].

variable name	units	mean	min	max
10 metre U wind component	$m\ s^{-1}$	1.05	-18.80	22.91
10 metre V wind component	$m\ s^{-1}$	0.57	-21.51	20.00
2 metre temperature	K	282.97	240.97	313.26
Leaf area index, high vegetation	$m^2\ m^{-2}$	1.84	-0.00	4.90
Leaf area index, low vegetation	$m^2\ m^{-2}$	2.24	-0.00	3.84
Low cloud cover	$(0\ -\ 1)$	0.38	0.00	1.00
Soil temperature level 1	K	283.12	257.68	313.64
Surface latent heat flux	$J\ m^{-2}$	-166716.23	-2203977.00	359411.00
Surface net thermal radiation	$J\ m^{-2}$	-192527.73	-664669.00	142945.02
Surface sensible heat flux	$J\ m^{-2}$	-41139.49	-1730277.12	826042.00
Total cloud cover	$(0\ -\ 1)$	0.66	0.00	1.00
Total column rain water	$kg\ m^{-2}$	0.01	-0.00	2.73
Total sky direct solar radiation at surface	$J\ m^{-2}$	282134.31	-0.12	3088320.00

Load Data

Besides weather data used to refine the forecasting results, also historic load data was needed. Therefore data has been retained from Open Power System Data[3]. Figure 4.3

[2]https://www.ecmwf.int/en/forecasts/datasets/browse-reanalysis-datasets
[3]https://data.open-power-system-data.org/time_series/

shows the distribution of loads over time with one point per day at 12 am UTC time. The color shows the mean temperature measured at 2 metres. The same data in a single plot is added in the Appendix as Figure A.1, it is split here in two plots for better distinction of single points.

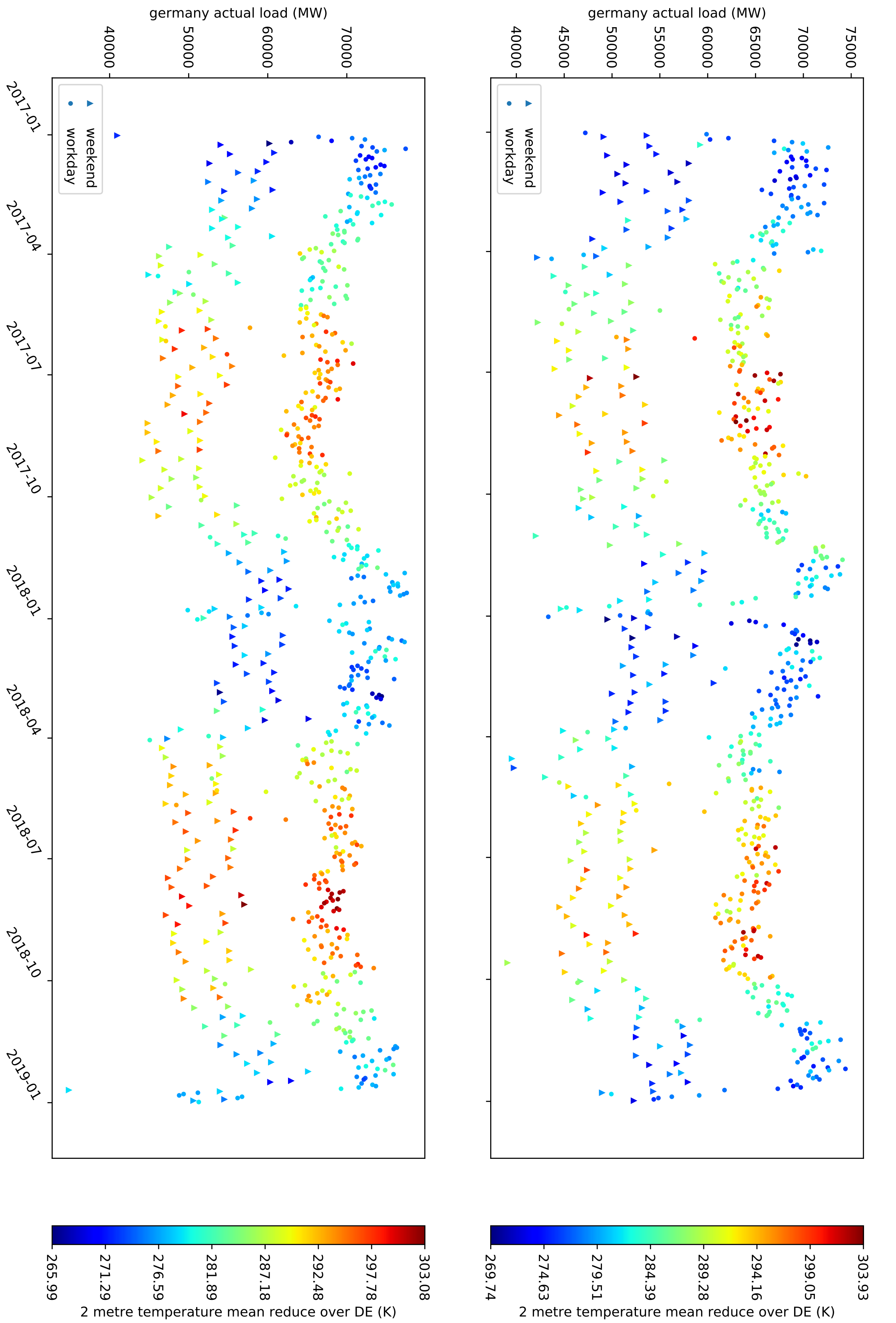

Figure 4.3. Load curve with mean of 2 metre measured temperature in Germany as color from 2015/1/1 to 2018/12/31 with one single point per day at 12 AM UTC time respectively.

Population Data

For further improvement and in order to filter important points in the grid, population data for Germany has been acquired from Eurostat[4]. The data is being visualized in Figure 4.4 with a logarithmic scale to improve visual distinction between the different regions which would be difficult for a non-logarithmic scale.

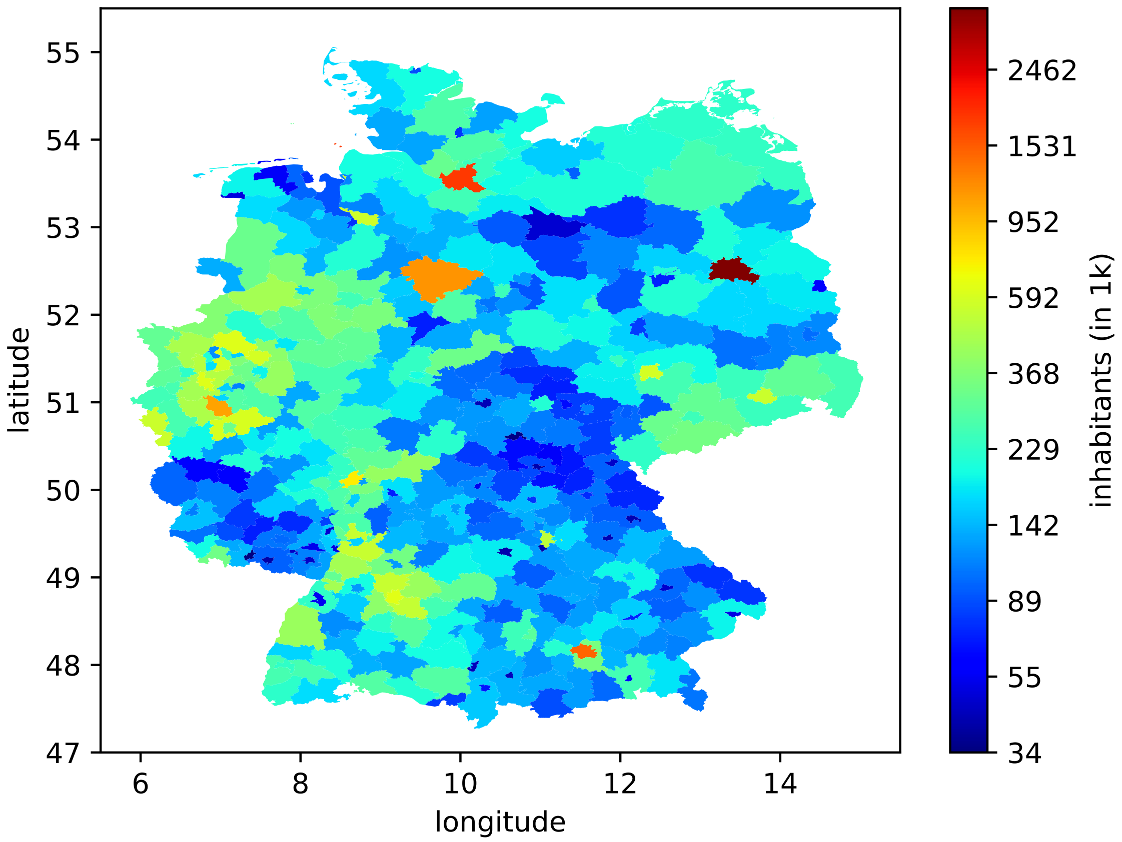

Figure 4.4. Population of Germany for each region respectively using a logarithmic scale for better distinction.

Data Preprocessing

As both, the load data and the weather data come from trustful sources, the need for data preprocessing is very limited. Nevertheless, there is still some missing or doubtful data. E. g. the values for the first months of the load data, that are located at the end of 2014, are much smaller than those after 2015. There is a possibility that this might be correct

[4] `https://ec.europa.eu/eurostat/data/database`

data, but to avoid uncertainties, only data from the beginning of 2015 to the end of 2018 is used. Also, there are some missing values that are linearly interpolated over time. The same about missing values applies to the weather data. Here it would also be possible to interpolate over locality, but for some time steps, there are no values at all, which is why temporal interpolation makes more sense in this case.

4.2. Implementation

This section will first outline which programming language is used and why. After that there are two points about the documentation and an additional comment about the process of implementation.

For the programming part, Python[5] 3.6+ has been chosen, as there is a variety of libraries to process all used file formats and because it tends to be a time saving language, also for visualization. Furthermore, Python has a highly vibrant and supportive community.

In regard to coding styles, especially when it comes to docstrings, the numpy conventions are used. This style is based on the reStructuredText[6] markup syntax. The three major points for this are first, that it is a popular and often used style, second, it is also a visually oriented style which means, that it is easy to read as can be seen in Figure 4.5 and last, it is supported by several automated documentation tools that create a PDF or HTML based documentation from existing source code with docstrings. For this purpose, the Sphinx[7] tool is used.

Considering the ARMA model, the statsmodels[9] package is used. During the process of implementation, it turned out that the forecasting functionality of this model does not behave correctly when using exogenous data. Therefore, this had to be implemented newly. At this point, I would like to thank Kaleb Phipps[10] who provided some very useful code, saving me a lot of time that would have otherwise been spent for debugging.

[5] https://www.python.org/
[6] http://docutils.sourceforge.net/rst.html
[7] http://www.sphinx-doc.org
[8] https://wingware.com/
[9] http://www.statsmodels.org/stable/index.html
[10] https://www.iai.kit.edu/2154_2880.php

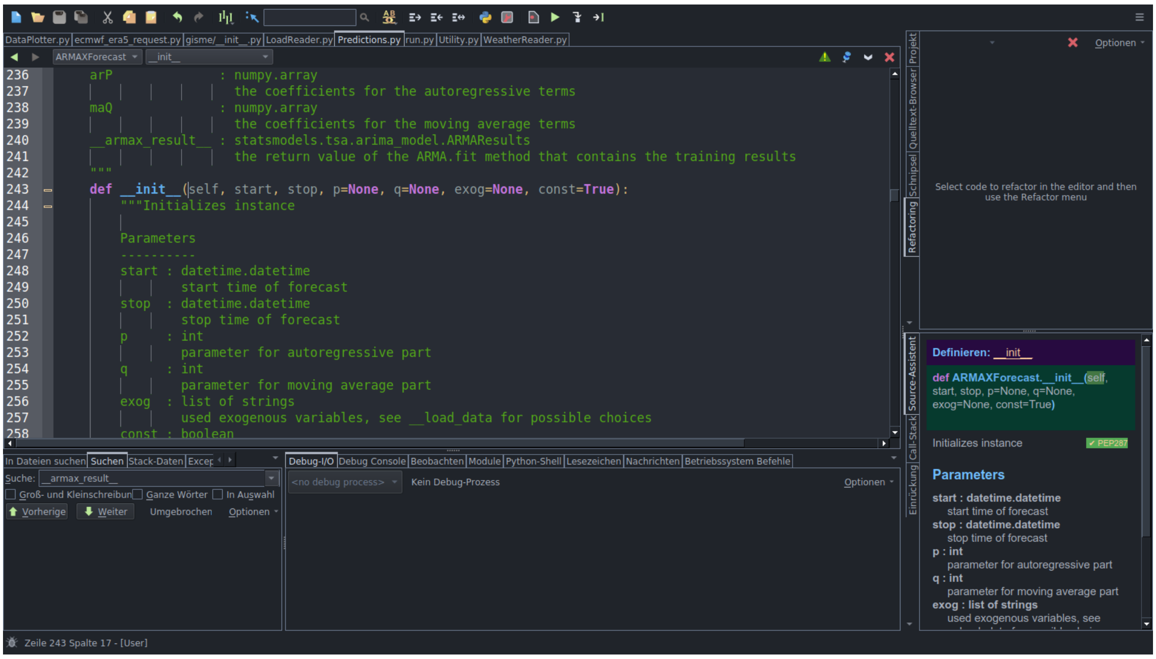

Figure 4.5. Python code and numpy docstring in the Wing IDE.[8].

4.3. Results

In this section, there are first two points about model selection. After that, the acquired results are presented.

Model Selection

The first steps for estimating reasonable values for p and q values of an ARMA often are the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). The respective plots in Figure 4.6 may suggest a p value between 3 and 6 and a q value between 2 and 4.

Information criteria are often used metrics for ARMA models when p and q are limited to a small range of numbers. Then all those models are fitted and the respective information criteria are computed. In general, as can be seen in Table 4.2, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan–Quinn Information Criterion (HQIC) get smaller for higher p and q values, which is the desired behaviour, as smaller values are better when comparing information criteria for different models that use the same data. However, for higher p and q values of 4 or 5, it can be observed, that the information criteria do not decrease considerably. Thus, it can be said, that appropriate p

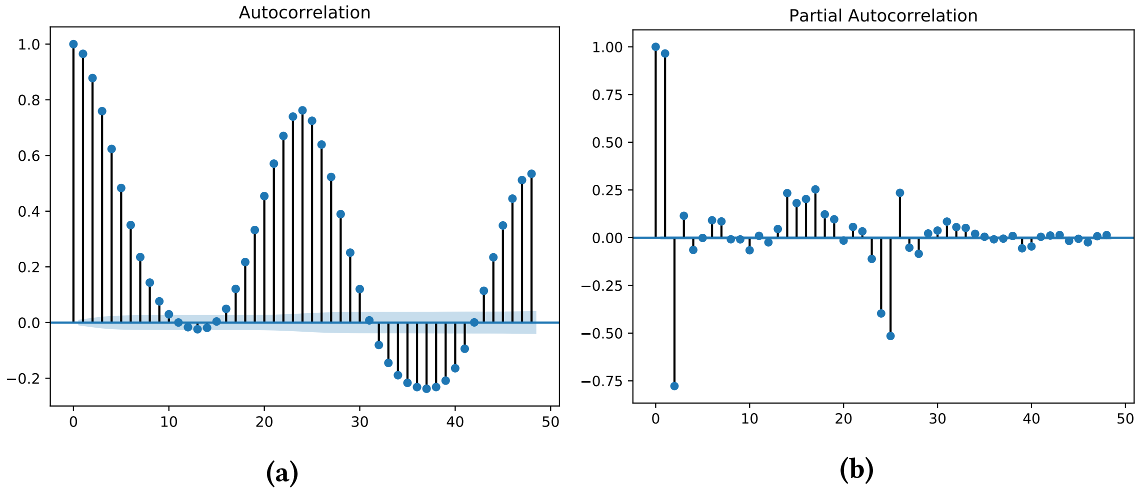

Figure 4.6. Autocorrelation (a) and Partial Autocorrelation (b) plots used to select the order of ARMA and ARMAX models.

and q values may be found between 2 and 4. The missing values in the table result from cases where some parameters could not be computed.

ARMAX

Figure 4.7, Figure 4.8 and Figure 4.9 show plots of one step ahead forecasts of ARMA(2,2) and ARMAX(2,2) models. This equals a one hour forecast due to the one hour resolution of the data. The training data has a time range from 2015/01/08 00:00 to 2017/12/31 00:00 and the forecast horizon covers the range from 2017/12/31 01:00 to 2018/12/31 00:00.

In Table 4.4, 168h shifted load means the time series, where the actual load data is shifted back in time by one week. The temp mean variable is the mean of the 2 metre temperature for each step in time, respectively. The data counter represents a counter, that counts the steps from 0 up to the length of the data. And last top 10 regions temp denotes the grid points of the 10 regions with the highest population, each grid point as a single time series.

As already can be estimated from the information criteria in Table 4.2, there is a huge difference between the ARMA(1,0) and the ARMA(2,2) model. This results in a difference of more than 1.3% of the MAPE and can be observed in Table 4.4. It can be seen, that of all exogenous variables used, including load data shifted by one week, has the greatest impact on the forecast error, causing these model to have the lowest error. In contrast to that, including the 10 regions with the highest population lead to a higher error. This

Table 4.2. Information Criteria for ARMA models without exogenous inputs using training data from 2015/01/01 to 2017/12/31 to compare for model selection with number of AR terms on one axis and number of MA terms on the other.

	MA(q) > AR(p) v	0	1	2	3	4	5
AIC	1	489489.23	472851.05	467543.07	466091.08	465622.11	464728.34
BIC		489513.76	472883.76	467583.96	466140.14	465679.35	464793.76
HQIC		489497.15	472861.61	467556.27	466106.92	465640.59	464749.47
AIC	2	464699.50	464253.96	464204.94	464190.14	464054.46	463634.12
BIC		464732.21	464294.85	464254.01	464247.39	464119.87	463707.72
HQIC		464710.06	464267.16	464220.79	464208.63	464075.58	463657.89
AIC	3	464324.69	464226.02	464202.92	463643.15	462483.99	462271.48
BIC		464365.58	464275.09	464260.16	463708.57	462557.59	462353.25
HQIC		464337.89	464241.87	464221.4	463664.27	462507.75	462297.88
AIC	4	464171.28	464173.25	462433.17	462985.37	462149.01	-
BIC		464220.35	464230.49	462498.59	463058.97	462230.78	
HQIC		464187.13	464191.73	462454.29	463009.14	462175.41	
AIC	5	464173.18	463588.38	462944.78	-	462138.70	-
BIC		464230.42	463653.80	463018.38		462228.65	
HQIC		464191.66	463609.50	462968.54		462167.74	

Table 4.3. The names of the ten regions with the highest population and the actual population for 2018, respectively.

region name	**population**
Berlin	3613495
Hamburg	1830584
München, Kreisfreie Stadt	1456039
Region Hannover	1152675
Köln, Kreisfreie Stadt	1080394
Frankfurt am Main, Kreisfreie Stadt	746878
Stuttgart, Stadtkreis	632743
Düsseldorf, Kreisfreie Stadt	617280
Recklinghausen	616824
Rhein-Sieg-Kreis	599056

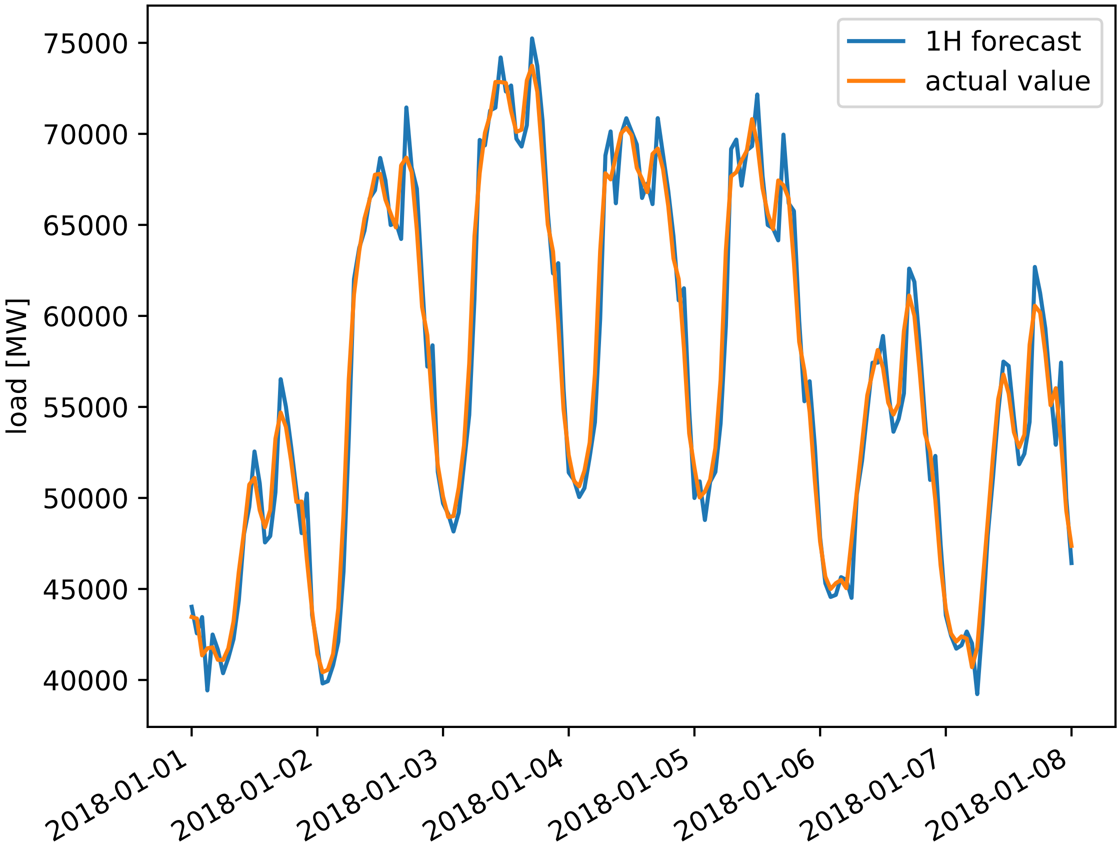

Figure 4.7. An example forecast from 2017/12/31 01:00 to 2018/01/08 00:00 for an ARMAX(2,2) using load data from 2015/01/08 00:00 to 2017/12/31 00:00 for training.

could result from the high correlation between neighbouring grid points in single regions. The respective regions with the highest population are listed in Table 4.3. A model for the 2 metre temperature with all grid points has been fitted too, but the number of 1435 exogenous variables implies a massive amount of computation time which is unpracticable for real-time forecasting. Furthermore, due to the high correlation between the different grid points, there is a high chance for errors during training. Also, a model with that many variables is prone to over-fitting and thus has a higher error on new data which can be seen in Figure 4.10. It is very interesting that including the mean temperature seems to result in a slightly better result than using any constellation of single grid points such as e. g. the 10 regions with the highest population.

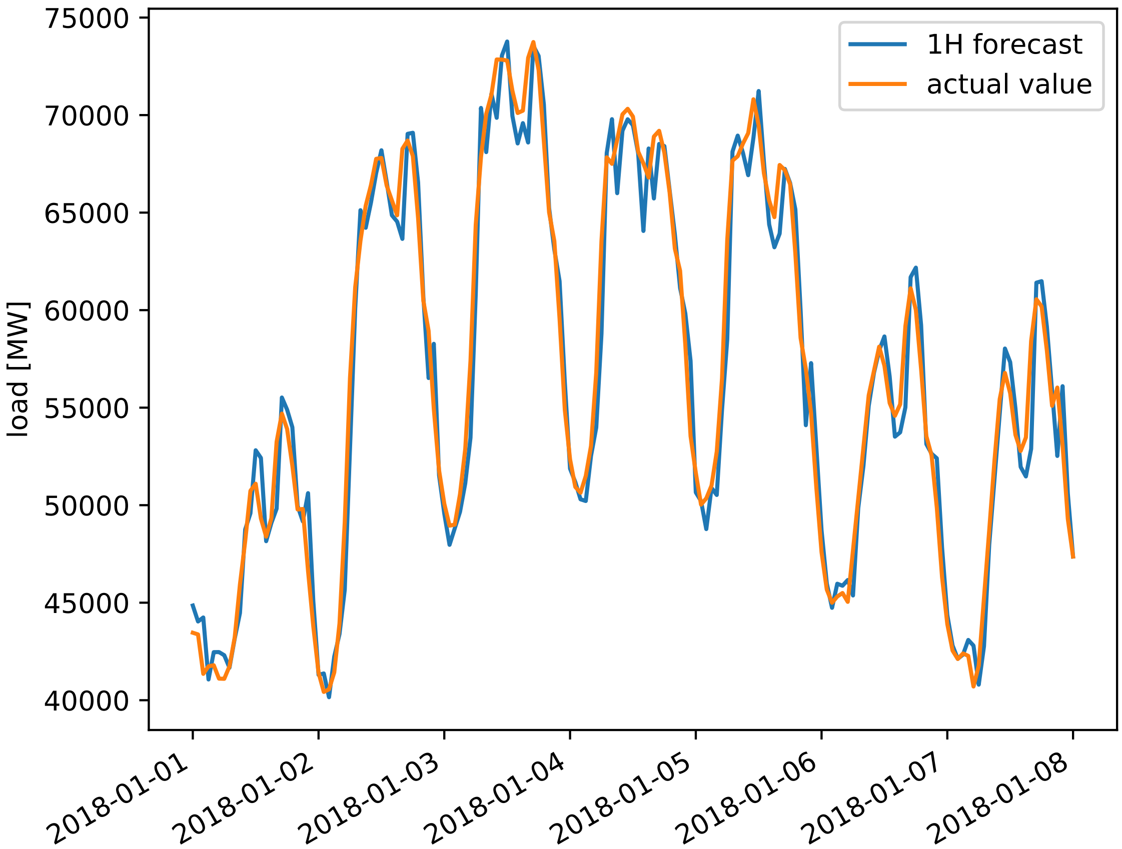

Figure 4.8. An example forecast from 2017/12/31 01:00 to 2018/01/08 00:00 for an ARMAX(2,2) using load data from 2015/01/08 00:00 to 2017/12/31 00:00 for training and the grid points of the ten regions with the highest population as exogenous input.

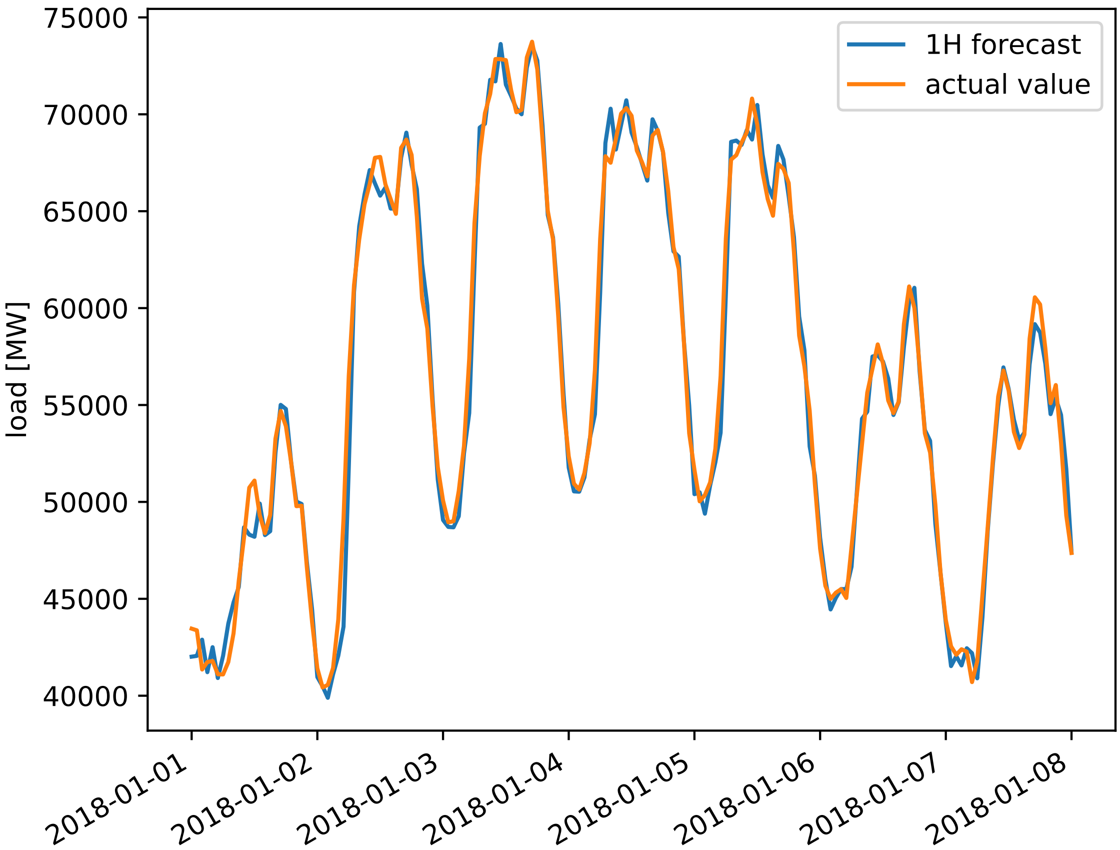

Figure 4.9. An example forecast from 2017/12/31 01:00 to 2018/01/08 00:00 for an ARMAX(2,2) using load data from 2015/01/08 00:00 to 2017/12/31 00:00 for training and the load data shifted back in time by one week for the same time range as exogenous input.

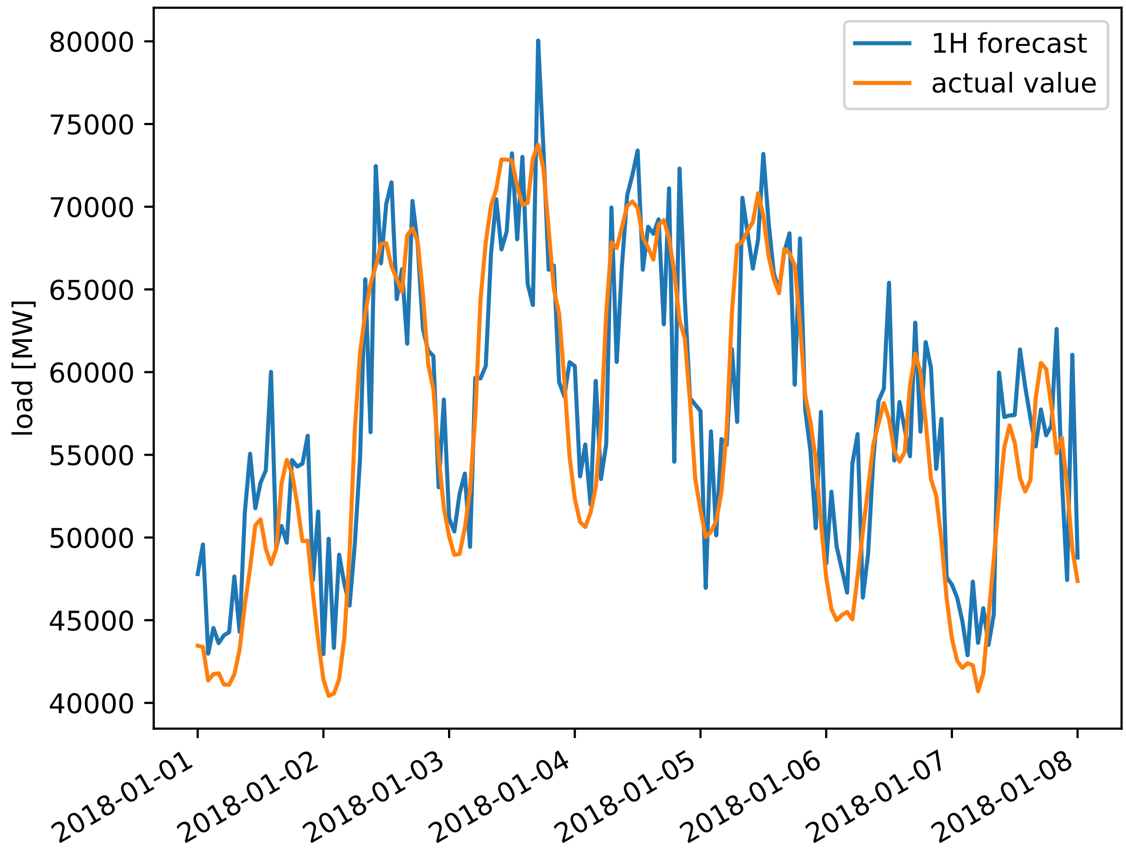

Figure 4.10. An example forecast from 2017/12/31 01:00 to 2018/01/08 00:00 for an ARMAX(1,0) using load data from 2017/01/01 00:00 to 2017/12/31 00:00 for training and all 1435 grid points of the 2 metre temperature variable as exogenous inputs.

Table 4.4. ARMA/ARMAX performance using different input data. For each error metric, the field with the best value is highlighted in green.

model	168h shifted load	temp mean	data counter	weekend	top 10 regions temp	RMSE	MAE	MPE	MAPE
ARMA(1,0)	✗	✗	✗	✗	✗	2621.940	1961.213	0.006	3.460
ARMA(2,2)	✗	✗	✗	✗	✗	1727.742	1216.559	0.220	2.114
ARMAX(2,2)	✓	✓	✗	✗	✗	1024.265	628.298	-0.020	1.103
ARMAX(2,2)	✓	✓	✓	✗	✗	1024.310	628.331	-0.006	1.103
ARMAX(2,2)	✓	✓	✗	✓	✗	1024.251	628.292	-0.020	1.103
ARMAX(2,2)	✓	✓	✓	✓	✗	1024.295	628.324	-0.006	1.103
ARMAX(2,2)	✓	✗	✗	✗	✗	1030.753	633.436	0.101	1.112
ARMAX(2,2)	✓	✗	✗	✓	✗	1030.756	633.433	0.101	1.112
ARMAX(2,2)	✓	✗	✓	✗	✗	1029.191	633.093	-0.165	1.114
ARMAX(2,2)	✓	✗	✓	✓	✗	1029.199	633.103	-0.165	1.114
ARMAX(2,2)	✓	✗	✗	✗	✓	1043.292	649.564	-0.018	1.144
ARMAX(2,2)	✓	✗	✓	✗	✓	1043.711	650.046	-0.012	1.145
ARMAX(2,2)	✓	✗	✓	✓	✓	1051.469	659.243	-0.012	1.164
ARMAX(2,2)	✓	✓	✗	✓	✓	1051.801	659.805	-0.021	1.165
ARMAX(2,2)	✓	✓	✓	✓	✓	1051.956	660.006	-0.004	1.165
ARMAX(2,2)	✓	✓	✗	✗	✓	1062.208	667.663	-0.011	1.179
ARMAX(2,2)	✓	✗	✓	✗	✗	1656.760	1178.435	0.021	2.043
ARMAX(2,2)	✗	✓	✗	✗	✗	1662.298	1179.187	0.185	2.039
ARMAX(2,2)	✗	✓	✗	✓	✗	1661.909	1179.681	0.187	2.040
ARMAX(2,2)	✗	✓	✓	✓	✗	1656.224	1179.021	0.022	2.044
ARMAX(2,2)	✗	✗	✓	✗	✗	1730.620	1226.238	-0.382	2.119
ARMAX(2,2)	✗	✗	✓	✓	✗	1730.620	1226.237	-0.382	2.119
ARMAX(2,2)	✗	✓	✓	✓	✓	1811.407	1321.255	0.043	2.315
ARMAX(2,2)	✗	✗	✓	✗	✓	2004.932	1444.396	-0.036	2.547
ARMAX(2,2)	✗	✓	✓	✗	✓	2032.248	1456.478	-0.025	2.570
ARMAX(2,2)	✗	✗	✗	✗	✓	2130.514	1508.978	0.211	2.663

5. Discussion

Considering the main research question of this thesis, whether grid-based weather information does have a benefit on energy forecasting, it is now possible to approach an answer. As already pointed out in Section 4.3, using the actual grid points themselves does not result in any improvement, but the results rather deteriorate. This seems to suggest that there is no benefit at all, using grid-based data for energy forecasting. However, grid-based data can be very useful when it is being conglomerated or compressed in a representative form, such as the mean over the longitude and latitude. This can also be observed in Section 4.3, where the averaged 2 metre temperature actually improves the forecast accuracy. The special upside in using grid-based data, such as the data from ECMWF, is, that it can be conglomerated over any desired locality as it is available for most locations. This allows it to be used for forecasts at arbitrary locations. Though, from the missing results to most of the given weather variables, it is can not be said which of them are most suitable to be used as inputs for energy forecasting. This provides an interesting subject for further research. But also other methods could be considered for further research, like filtering grid points by economic activity. Another possibility would be to use completely different models such as Recurrent Neural Networks (RNN), which are particularly suitable for time series forecasting.

6. Conclusion

This thesis addressed the subject of whether grid-based weather information provides a benefit in energy forecasting or not. It may not appear entirely clear that the answer is yes, as there definitely is a benefit, even though not as expected. The behaviour that has been expected, is, that e. g. filtering the most populated regions and using those grid points was to improve the forecast result, but it actually worsened it. This could be observed for a forecast using the huge number of 1435 exogenous variables, one for each grid point. It turned out, that this does not only highly comprise the computation time, but also has a very negative impact on the accuracy of the forecast. Still, there is an improvement for using e. g. averaged temperature data, which is a noteworthy benefit of grid-based data, as the average temperature can be computed for any desired composition of grid points. It also needs to be mentioned, that this subject has not yet been directly addressed by any of the found related works. Previous works mainly focused on the actual forecast and how to improve it, rather than which sort of data actually provides beneficial behaviour in terms of forecasting. Further research is needed in order to figure out how grid-based data can be compressed optimally, so that the number of variables is limited to a small enough amount or generalizes well enough to avoid over-fitting. Further, forecasts with an increased time scope could be used, to evaluate the temporal range for which the current weather improves forecasts of the short-term future energy demand. In the end, there is an additional assumption made, which is, that in the near future, weather may have an even higher influence on energy consumption due to the current energy transition and possibly resulting outcomes such as DSM. This assumption emphasizes the importance of this research topic and also the practical usefulness of the results of this thesis.

Bibliography

Aguiar, L. M., B. Pereira, P. Lauret, F. Díaz, and M. David (2016). *Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting*. In: *Renewable Energy*, Vol. 97, pp. 599–610.

Alessandrini, S., L. Delle Monache, S. Sperati, and J. N. Nissen (2015). *A novel application of an analog ensemble for short-term wind power forecasting*. In: *Renewable Energy*, Vol. 76, pp. 768–781.

Bofinger, S. and G. Heilscher (2014). *Solar electricity forecast : Approaches and first results.* In: No. January 2006.

Davò, F., S. Alessandrini, S. Sperati, L. Delle Monache, D. Airoldi, and M. T. Vespucci (2016). *Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting*. In: *Solar Energy*, Vol. 134, pp. 327–338.

De Felice, M., A. Alessandri, and F. Catalano (2015). *Seasonal climate forecasts for medium-term electricity demand forecasting*. In: *Applied Energy*, Vol. 137, pp. 435–444.

Diagne, M., M. David, P. Lauret, J. Boland, and N. Schmutz (2013). *Review of solar irradiance forecasting methods and a proposition for small-scale insular grids.* In: *Renewable and Sustainable Energy Reviews*, Vol. 27, pp. 65–76.

Fairley, I., H. C. Smith, B. Robertson, M. Abusara, and I. Masters (2017). *Spatio-temporal variation in wave power and implications for electricity supply.* In: *Renewable Energy*, Vol. 114, pp. 154–165.

Haben, S., G. Giasemidis, F. Ziel, and S. Arora (2018). *Short term load forecasting and the effect of temperature at the low voltage level.* In: *International Journal of Forecasting*, No. xxxx.

Hyndman, R. and G. Athanasopoulos (2018). *Forecasting: principles and practice, 2nd edition.* OTexts: Melbourne, Australia. URL: `https://otexts.com/fpp2` (visited on 07/01/2019).

Kamińska-Chuchmala, A. (2014). *Spatial internet traffic load forecasting with using estimation method.* In: *Procedia Computer Science*, Vol. 35, No. C, pp. 290–298.

Li, Y., V. G. Agelidis, and Y. Shrivastava (2009). *Wind-solar resource complementarity and its combined correlation with electricity load demand.* In: *2009 4th IEEE Conference on Industrial Electronics and Applications, ICIEA 2009*, pp. 3623–3628.

Ludwig, N., S. Feuerriegel, and D. Neumann (2015). *Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests.* In: *Journal of Decision Systems*, Vol. 24, No. 1, pp. 19–36.

Morley, J., K. Widdicks, and M. Hazas (2018). *Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption.* In: *Energy Research and Social Science*, Vol. 38, No. August 2017, pp. 128–137.

Salcedo-Sanz, S., R. C. Deo, L. Cornejo-Bueno, C. Camacho-Gómez, and S. Ghimire (2018). *An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia.* In: *Applied Energy*, Vol. 209, No. July 2017, pp. 79–94.

Sperati, S., S. Alessandrini, and L. Delle Monache (2016). *An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting.* In: *Solar Energy*, Vol. 133, pp. 437–450.

A. Appendix

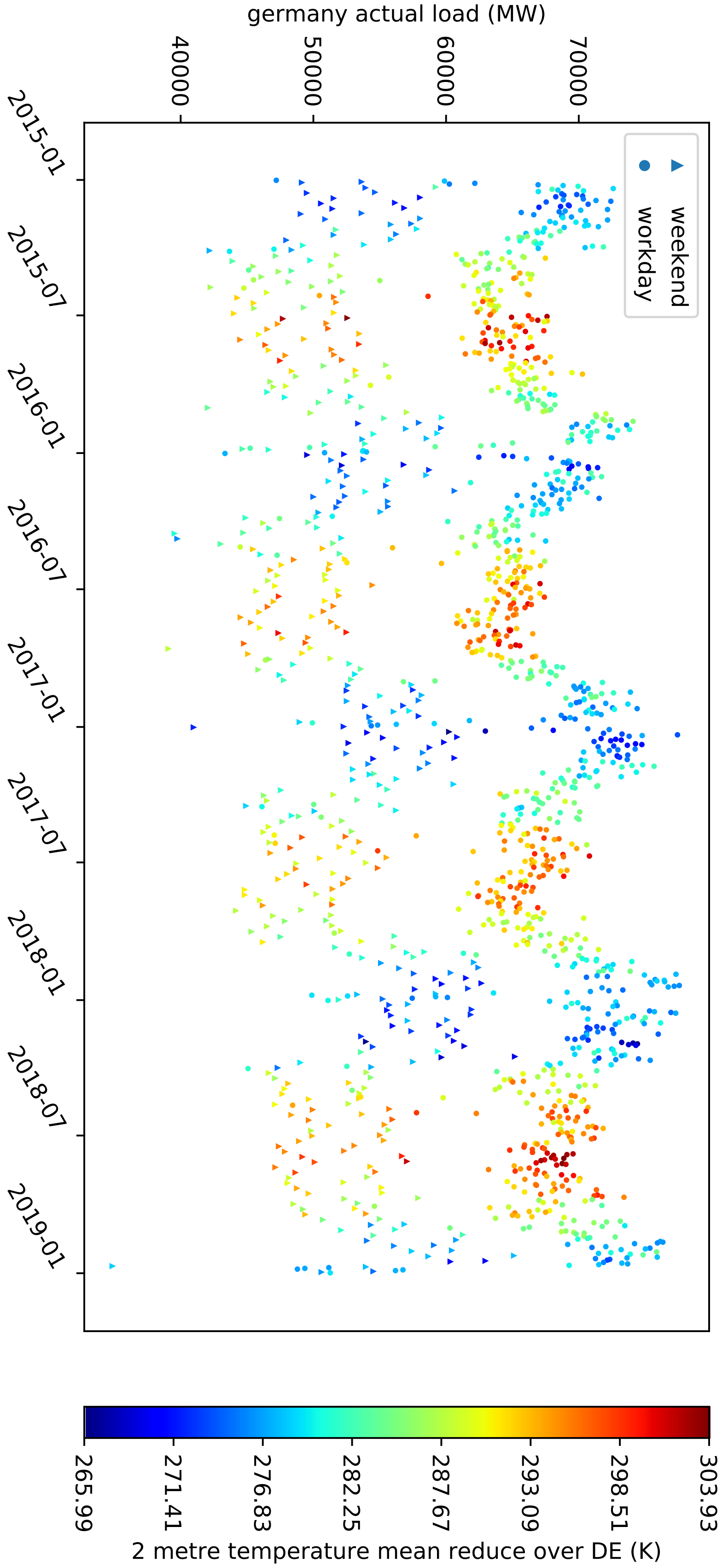

Figure A.1. Load curve with mean of 2 meter height measured temperature in germany as color from 2015/1/1 to 2018/12/31 with one single point per day at 12am utc time respectively.