A Statistical and Machine Learning Approach to Air Pollution Forecasts

Simon Carlén

Degree project, 15 credits
Computer and Systems Sciences
Degree project at the master level
Spring term 2022
Supervisor: Sindri Magnússon
Co-supervisor: Ali Beikmohammadi

Stockholm
University

Abstract

Synopsis

Background

Problem

Research Question

Method

Result

Discussion

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

1. Introduction

1.1 Background

Outdoor air pollution is a major global environmental issue, linked to several serious health conditions, and causing millions of premature deaths every year [1]. Some principal air pollutants damaging to health include gaseous substances such as nitrogen oxides (NO_x), ground-level ozone (O_3), sulphur dioxide (SO_2), and carbon monoxide (CO), but also atmospheric aerosol particles such as PM_{10} and $PM_{2.5}$ [2]. In Stockholm, traffic is a major source of local air pollution, and though air quality is generally good, some streets experience short episodes with severe pollution levels, especially during winter and spring [3].

To protect public health, urban air is normally monitored. In addition to monitoring, forecasts of air quality (both hourly and daily) can be critical to regulatory authorities, and in general, there are two approaches to this; with mechanistic models or statistical and/or machine learning models. [4–6]. With mechanistic models, the processes governing the evolution of air pollution is modeled mathematically, whereas statistical and machine learning models are more data-driven [6].

From a statistical perspective, predicting air pollution is a time series regression problem, and there are many different regression techniques for forecasting and time series analysis [6]. These techniques can vary in complexity, from more simple linear models to deep neural networks capable of finding complex non-linear relationships in the data [6, 7]. Nonetheless, one of the main challenges with air pollution is that there are dependencies over both space and time (i.e., the data is spatio-temporal), and simpler models may not capture these dependencies [5]. Recent advances in machine learning however have shown promising results when it comes to air quality forecasts, especially deep neural networks [5, 6].

1.2 Research problem

Forecasts, be it for weather, stock returns, or future pandemics, are always associated with uncertainty and errors. Erroneous predictions made by existing air pollution forecasting systems, both mechanistic and statistical and/or machine learning-based, can be attributed to many causes. In the case of mechanistic models, there can be insufficient information about the necessary parameters needed for simulation and modeling [6]. For statistical and/or machine learning methods, too simplistic models, lack of data, irrelevant input features, overfitting, etc., can limit prediction accuracy [6].

Nevertheless, atmospheric pollution is a very complex phenomenon depending on a multitude of factors across both space and time. Hence, the research problem addressed in this work is:

- To capture and model the complex dynamics of air pollution with modern machine learning methods, in particular deep learning.

1.3 Research question

From a forecasting perspective, of special interest are episodes when pollution levels peak. Generally, this is also when existing forecasting models tend to give the largest prediction errors [6]. Therefore, the research question this thesis tries to answer is:

- How can machine learning, in particular deep learning, be used to forecast air pollution levels and pollution peaks?

1.4 Delimitations

In this work, historical air pollution and weather data is used. Therefore, the models cannot be tested in "operational mode", i.e., with real-time data to make predictions. Moreover, when forecasting air pollution (hours or days ahead), weather forecasts are often utilized in addition to monitoring data to improve forecast results. Again, with historical data, incorporating weather forecasts is not possible, and consequently the models make predictions based only on multivariate time-series of past observations. This also puts a limit on the time-horizon for the predictions, since without weather forecasts, medium to long-term forecasts of air quality would have large uncertainties.

pollution background levels...

2. Extended Background

2.1 Ambient air pollution

Ambient air pollution is one of the greatest environmental and health concerns of the modern world. Worldwide, poor air quality causes millions of premature deaths every year and is linked to several adverse health effects such as respiratory problems, cardiovascular disease, and cancer [1]. In addition to health risks, the global economic impacts are substantial due to lost labor productivity, increased health care costs, reduced crop yields, etc. [8]. Outdoor air pollution has become a ubiquitous problem, affecting both cities and rural areas, and it is estimated that about 90% of the world's population are living in regions where air pollution levels exceed guidelines set by the World Health Organization [1].

2.1.1 Principal air pollutants

In densely populated urban areas, air pollution levels can periodically be severe, and with an accelerating urbanization, it has become imperative for regulatory authorities to closely monitor city air and try to mitigate the harmful effects of pollution. Commonly monitored substances include sulphur dioxide (SO_2), nitrogen oxides (NO_x, i.e., NO and NO_2), carbon monoxide (CO), ground-level ozone (O_3), volatile organic compounds (VOCs), and particulate matter (PM) [2].

Vehicular traffic is a major source of the gaseous pollutants NO_x, SO_2, CO, and VOCs, but certain industrial processes also contribute to emissions [2]. Ground-level O_3 (also a gas) is a so-called secondary pollutant that forms when NO_x and VOCs react on sunny days with little wind [2].

PM – the group of pollutants being the focus of this work – are atmospheric aerosol particles (i.e., particles suspended in the air). They have diverse origins, both natural and anthropogenic, and a complex chemical composition consisting of both solid and liquid species [9]. Some important sources of PM are forest fires, volcanic eruptions, sand/dust storms, sea spray, vehicular traffic, certain industrial processes, construction sites, and domestic combustion [9,10]. When entering the atmosphere directly by these routes, one denotes the PM as primary. However, PM can also be formed by the oxidation of gases such as SO_2, NO_x, and VOCs (followed by a complex chemical reaction process), in which case the PM is said to be secondary [9]. PM is also categorized by particle size (or more specifically, the aerodynamic diameter), and particles measuring smaller than 2.5 µm and 10 µm are denoted as $PM_{2.5}$ and PM_{10}, respectively [9].

Both PM_{10} and $PM_{2.5}$ can travel long distances from point sources (though $PM_{2.5}$ has a longer residence time in the atmosphere than PM_{10}), and local pollution can be affected by regional background levels [3, 9]. PM levels are also dependent on weather conditions [9]. For example, temperature and solar radiation are related to the formation of secondary PM, and PM emission from roads, tires, brake wear, etc., can be affected by precipitation and humidity [3, 6]. Both PM_{10} and $PM_{2.5}$ are hazardous and cause a wide range of health problems, though $PM_{2.5}$ can more easily penetrate the lungs [9]. In the European Union, annual mean limits are set to 40 $\mu g/m^3$ for PM_{10} and 20 $\mu g/m^3$ for $PM_{2.5}$ [11].

2.1.2 Ambient air pollution in Stockholm

In the city of Stockholm, environmental air quality standards are usually met, though some streets experience occasional episodes with severe pollution levels (e.g. Hornsgatan is one such street) [12]. Since Stockholm has centralized district heating and few industries, the major source of local CO, NO_x, and PM pollution is vehicular traffic [3, 12]. Mechanical wear by studded tires on asphalt and the wearing of brakes and tiers in motor vehicles contribute substantially to local levels of both PM_{10} and $PM_{2.5}$. For $PM_{2.5}$ however, contribution from sources outside of Stockholm is also significant [12]. Emission of SO_2 can come from the energy sector and waterborne transport, though local levels are also affected by outside sources. For O_3, long-range transport from mainland Europe is the single-most important factor contributing to locally measured levels [12].

The air in Stockholm County is monitored by Stockholms Luft- och Bulleranalys (SLB-analys), a unit in the Environment and Health Administration (EHA) of the city of Stockholm. SLB-analys are responsible for a number of monitoring stations measuring several air pollutants and some meteorological parameters in the Stockholm region, as well as a few stations outside of Stockholm [13]. In addition to monitoring the air, SLB-analys also model and forecast air pollution levels for the Stockholm metropolitan area, and their forecasts are available through a smartphone application, called "Luft i Stockholm" [3].

2.2 Forecasting air pollution

Having the possibility to forecast air pollution levels hours or days ahead can be extremely valuable to regulatory authorities in order to protect public health, and vulnerable groups in particular. In general, there are two broad categories of models for such forecasts; mechanistic models, and statistical and/or machine learning models [4]. This work is concerned with the latter type, and in the sections below a review follows. The mathematical and statistical theory behind many of the models is quite extensive [7, 14–16], but relevant theory will be covered briefly.

2.2.1 Forecasting as a regression problem

While mechanistic models are based on mathematical modelling of atmospheric processes along with other factors governing the distribution of air pollution (such as emission source characteristics, physico-chemical properties of pollutants, terrain and building design, etc.) statistical and/or machine learning models are entirely data-driven, being derived directly from measurements on the variables of interest [4]. From a statistical (or machine learning) perspective, forecasting air pollution can be viewed as a regression problem, in which a function f, mapping input data to a numerical output, is being approximated (or learned) from a training set of labeled input-output examples [16]. Learning the function f amounts to finding a set of parameters (or weights) for the model, which in the case of a simpler regression technique can be only a handful, but possibly millions if a deep neural network is used [16]. Generally in regression, the weights are learned by minimizing a cost function

$$J(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \big(\hat{y}_i - y_i\big)^2 \tag{2.2.1}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated model parameters ($\hat{\beta}_0$, $\hat{\beta}_1, ...,$ $\hat{\beta}_n$), $\hat{y}_i$ is a prediction and y_i is a training data value [16]. In Eq. 2.2.1 the squared error loss is used as loss function, and the cost is simply the loss averaged over the training data.[1] Depending on the model, minimizing $J(\hat{\boldsymbol{\beta}})$ is approached differently, as explained further in the sections below.

2.2.2 Linear models

From the wealth of available regression techniques, multiple linear regression (MLR) has been extensively used to forecast and model air pollution [6]. If none of the basic model assumptions are violated (i.e., linearity, independence, normality, and constant variance), MLR is often a straightforward method, especially for data with no temporal dependencies (so-called cross-sectional data). However, for time series data, the assumption of independent errors is often not approptiate [15].

If fitting a MLR model to time series data, successive errors will typically be correlated (often referred to as autocorrelation), and this will cause several problems with the model if the correlation is not accounted for [15]. To this end, adjustments to the MLR model can be made, some of which require other parameter estimation techniques than the usual least squares method (see below). However, a simple and commonly used procedure to get rid of the autocorrelation is to include one or more lagged values of the response variable as predictors. For example, if the value of the response variable at lag one is included, the MLR model will have the form

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{2,t} + ... + \beta_k x_{k,t} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{2.2.2}$$

where the error term $\varepsilon_t \sim N(0, \sigma^2)$, and t denotes time steps [15]. The model in Eq. 2.2.2 can be fit with the method of least squares, which in linear regression is the standard way of finding parameters so that $J(\hat{\boldsymbol{\beta}})$ is minimized [16]. This is done by solving the so-called normal equations

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \tag{2.2.3}$$

and the least squares estimates of the model parameters will then be given by Eq. 2.2.4 below (provided that the inverse of $\boldsymbol{X}^T\boldsymbol{X}$ exists) [16].

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.2.4}$$

Careful variable selection in regression is crucial as it can influence the performance of a model. In situations with several variables, one is often concerned with finding an optimal "subset" of predictors, where multicollinearity should also not be an issue [17].

[1] What is meant by cost and loss functions can vary slightly in the literature, but in this work, the terminology of Lindholm et al. [16] is adopted.

To this end, variable selection techniques based on optimizing a criterion like the Akaike or Bayes information criterion are common, and typically multicollinearity is also tested for [17]. However, if one is reluctant to exclude variables, but multicollinearity still might be an issue, regularized versions of MLR can be used [16, 17].

Two common techniques are L_1 and L_2 regularization, in which an extra so-called "penalty" term is added to the cost function to shrink the estimated model parameters. In L_2 regularization (also called ridge regression), the parameters will be pushed towards small values, whereas in L_1 regularization (or lasso regression), some parameters will be driven to zero. The penalty terms for ridge and lasso regression are, respectively,

$$\lambda \sum_{j=1}^{k} \beta_j^2 \quad \text{and} \quad \lambda \sum_{j=1}^{k} |\beta_j|$$

where λ is a parameter controlling the shrinkage [16]. For ridge regression, the parameter estimates can be found by solving a modified version of Eq. 2.2.3, while for lasso regression, no such analytical solution exists, and numerical optimization techniques have to be used instead [16]. By shrinking the parameters, ridge and lasso regression works as a variable selection method, while also preventing overfitting when used in more complex regression models [16].

The extensive use of MLR for air pollution forecasts is many times motivated by its simplicity and straightforward implementation [6]. Another advantage is interpretability; for example, inference can be made on all input variables, allowing one to investigate their individual importance [17]. However, the assumption of linearity might not always hold, and rather large prediction errors have been observed at times of pollution peaks [6]. Moreover, with data from several (but nearby) monitoring stations, collinearity can be an issue, which is why ridge or lasso regression are popular alternatives to the more classical non-regularized MLR model [18].

2.2.3 Extensions of the linear model

More versatile and flexible regression models generally produce better forecasting results than linear techniques [6]. Some examples include regression trees, generalized additive models, and support vector machines (SVM) [6, 18]. These models can handle more complex non-linear input-output relationships, and especially SVM has been successfully applied for PM_{10} prediction, sometimes with better results than artificial neural networks [6].

Artificial neural networks (ANNs), in particular the multilayer perceptron (MLP), have also been extensively used as a forecasting technique [6]. ANNs are flexible models able to handle non-linear input-output relationships, however, over-fitting can be an issue, especially with high-dimensional input and if training data is limited [6, 18].

The MLP is a so-called feedforward neural network, in which a set of input data is taken and passed through several "hidden" layers made up of neurons (also called units), before an output is produced [7]. Deep neural networks can have many such layers (hence the term "deep" [19]), and each layer can have hundreds of units. Every layer produces a slightly more abstract representation of its input by non-linear transformations, and with several such transformations, complex relationships in the data can be learned [7].

Many other deep learning architectures than the MLP exist, such as convolutional neural networks (CNNs), or recurrent neural networks (RNNs). CNNs are commonly used for image recognition while RNNs (and variants thereof) normally are applied to sequential data.

2.2.4 Variable selection

In any regression problem, variable/feature selection is crucial as it can influence the performance of a model. With regards to PM, as pointed out in section 2.1.1, weather conditions can greatly affect pollution levels, and therefore meteorological data can be utilized to improve forecasts [6].

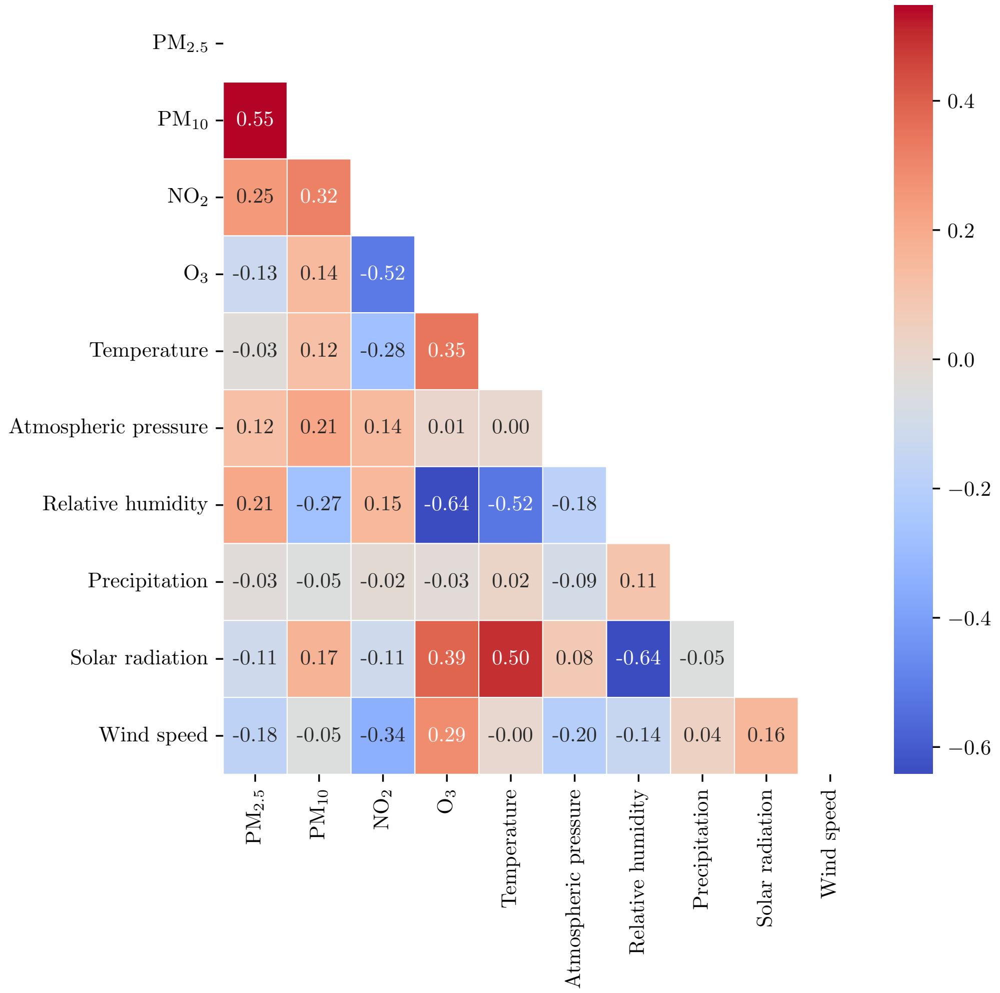

Figure 2.1: Pairwise correlations between air pollutants and some meteorological variables.

Additional variables can also be included in PM forecasts. For example, motor traffic data such as travel speeds, traffic flow and intensity, etc., can be utilized [6]. Data on other pollutants can also be important, especially SO_2 and NO_x as they are involved in the formation of secondary PM [20]. Moreover, if forecasts focus solely on PM_{10} (as in this work), data on $PM_{2.5}$ can further improve the results [20]. Temporal variables such as time of the day and time of year are also useful since daily and seasonal variation of PM pollution is important [6, 9].

In Figure 2.1 where pairwise correlations between a few meteorological variables and PM_{10} at different stations in the Stockholm region are given (see Table 3.1 for details about the different monitoring stations), it can be inferred that PM_{10} correlate negatively with humidity, but positively with atmospheric pressure and solar radiation. It can also be seen that PM_{10} levels are strongly correlated among some stations.

In this work, in addition to PM_{10} data, meteorological data as well as data on $PM_{2.5}$ and NO_x were utilized. Some features used as input to the models were also derived. A more detailed description of the variables and their preprocessing is given in section 3.1.

2.3 Summary and motivation for this work

3. Methodology

The major steps of the implemented workflow is shown in Figure 3.1. Historical air pollution data from several monitoring stations, together with meteorological data from one station, was retrieved, preprocessed (with some features engineered), and divided into data windows. Three deep learning models (feed forward neural network, RNN, and LSTM) were trained and tested for short-term predictions (one hour ahead) of PM_{10} for one station at Torkel Knutssongatan (measuring urban background levels, see Table 3.1). As baseline models for comparison, multiple linear regression and ARIMA were used. Detailed descriptions of each step in the process are given in subsequent sections.

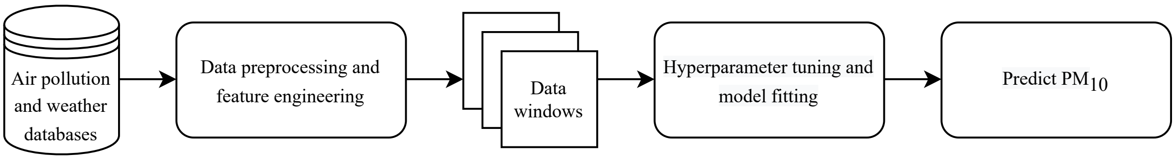
Air pollution and weather databases
Data preprocessing and feature engineering
Data windows
Hyperparameter tuning and model fitting
Predict PM10

Figure 3.1: Implemented workflow.

3.1 Data retrieval and preprocessing

3.1.1 Data sources

Air pollution data was retrieved from the Swedish Meteorological and Hydrological Institute's (SMHI) centralized database for air quality measurements [21]. This data is part of the national and regional environmental monitoring of Sweden, a program coordinated and funded by the Swedish Environmental Protection Agency (Swedish EPA) and the Swedish Agency for Marine and Water Management. There are in total ten different program areas, of which air is one, and all data are licensed under CC0 and therefore freely accessible to the public [22]. For the national air monitoring (under Swedish EPA's responsibility), SMHI acts as national data host and stores (quality checked) historical data reported on a yearly basis from municipalities in Sweden [21].

In Stockholm County, there are 18 stationary monitoring sites, and initially, data from each station was considered. However, many stations had irregular data series, and not all of them measure the same set of parameters. Due to this, data from seven sites with hourly measurements of PM_{10} and $PM_{2.5}$ ($\mu g/m^3$) for the period 2016-01-01 (01:00 am) to 2020-09-16 (01:00 am) was chosen, giving in total 41,281 data points (per parameter). In six of these stations, hourly data for NO_x was also available. Air pollution monitoring can be classified by the area surrounding the station (rural, rural-regional, rural-remote, suburban, and urban), and by the predominant emission sources (background, industrial, or traffic) [21]. The chosen stations included data from both traffic and background monitoring, in urban as well as rural-regional areas. Information about the stations from which data was used is summarized in Table 3.1.

As described in section 2.1.2, SLB-analys also monitor a number of weather parameters, and hourly measurements of temperature (in °C), precipitation (mm), atmospheric pressure (hPa), relative humidity (as %), solar radiation (W/m^2), wind speed (m/s), and wind direction (in degrees) was included from the station at which the PM$_{10}$ predictions were made (Torkel Knutssongatan). The meteorological data was downloaded from SLB-analys' webpage [23].

Table 3.1: Summary of monitoring stations in Stockholm County.

Station	Station code	Longitude	Latitude	Classification	Parameters
Norrtälje, Norr Malma	18643	18.631313	59.832382	Rural-regional background	PM$_{10}$, PM$_{2.5}$, NO$_x$
Sollentuna, E4 Eriksbergsskolan	34399	17.957651	59.410175	Urban background	PM$_{10}$, PM$_{2.5}$
Sollentuna, E4 Häggvik	20415	17.922358	59.443535	Urban traffic	PM$_{10}$, PM$_{2.5}$, NO$_x$
Stockholm, E4/E20 Lilla Essingen	18644	18.00439	59.325527	Urban traffic	PM$_{10}$, PM$_{2.5}$, NO$_x$
Stockholm, Hornsgatan 108 Gata	8780	18.04866	59.317223	Urban traffic	PM$_{10}$, PM$_{2.5}$, NO$_x$
Stockholm, Sveavägen 59 Gata	8779	8.058254	59.340828	Urban traffic	PM$_{10}$, PM$_{2.5}$, NO$_x$
Stockholm, Torkel Knutssongatan	8781	18.057808	59.316006	Urban background	PM$_{10}$, PM$_{2.5}$, NO$_x$, Meteorological parameters

3.1.2 Data preprocessing

Time series plots of the raw data for PM$_{10}$ and PM$_{2.5}$ is shown in Figure 3.4. As can be seen e.g. in plot (e) and (f), some stations had short periods with missing data, and linear interpolation was used to fill in the missing data points. Moreover, all data was min-max normalized (i.e., scaled to the range [0,1]) before use in any of the models.

Feature engineering From the meteorological data, wind vectors (u and v) were derived from wind direction and wind speed, as wind vectors are more suitable model inputs [24]. After converting wind direction values to radians, u and v were obtained in the following way

$$u = ws * cos(\theta)$$

$$v = ws * sin(\theta)$$

where ws denote wind speed and θ is the wind direction (in radians). From Figure **??**, yearly periodicity in the data can be seen, where levels tend to be higher during spring. Daily periodicity is also expected since traffic intensity varies throughout the day. The meteorological parameters such as temperature, solar radiation, etc. also have yearly and daily periodicity. To account for this, time was converted to sine and cosine signals, both for year and day, as a way to model these cyclic features [18]. Sine and cosine for day were calculated in the following way

$$sine\ day = \frac{sin\left(timestamp \cdot \frac{2\pi}{86{,}400}\right) + 1}{2}$$

$$cosine\ day = \frac{cos\Big(timestamp \cdot \frac{2\pi}{86{,}400}\Big) + 1}{2}$$

where *timestamp* is in seconds (and with 86,400 seconds in 24 hours, dividing by this term is necessary). The calculations were done similarly for year except for the term in the denominator which instead was set to seconds per year ($365.25 \cdot 86,400$). The transformations were done so that the sine and cosine functions oscillate between zero and one. The resulting transforms for day (in a 24 hour time window) and year (time window of one year) are shown in Figure 3.2a and Figure 3.2b, respectively.

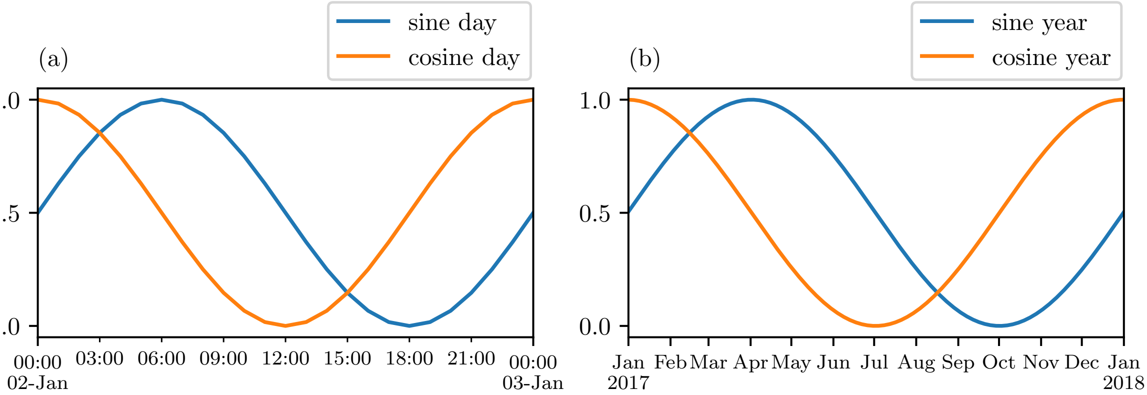

Figure 3.2: (a) Sine and cosine of day, and (b) sine and cosine of year.

Sliding windows Sliding windows from the data were also created. The sliding window approach is used for time-series forecasting where windows (or sequences of certain lengths, also called frames) are extracted from the input data [20, 25]. In each window, there are two ”sub-windows”; the input window and the target window, and the target window is offset by some amount of time from the input window. For example, as shown in Figure 3.3, the total window length is nine time steps, and the first eight time steps is the input window used to predict the target window (in this case having a length of one) one time step in the future. After extracting a data sequence, the window slides to the right one (or more) steps and extracts the next sequence. This is continued until time step n at which point all the data have been processed. In this work, input windows of different lengths were tested to make short-term predictions for a target window with a length of one (more details are given in section 3.2 below).

Train-test split Lastly, the data was split into training, validation, and test sets, where the validation set was used for hyperparameter optimization. The test set was taken as the most recent year of data (from 2019-09-16 to 2020-09-16), the validation set was taken as the year prior to the test data (2018-09-16 to 2019-09-16), and the remaining data was used for training (2016-01-01 to 2018-09-16). This split is motivated by the fact that the data is in the form of time-series, where each observation has a specific time-stamp and where successive observations are (in this case) positively autocorrelated [26].

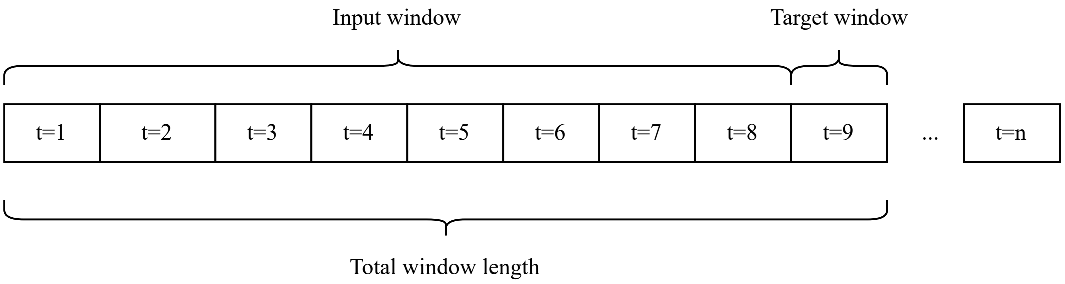

Figure 3.3: Sliding window approach for time-series data.

3.2 Hyperparameter tuning and search

The Keras Tuner library [27] was used to find the best set of hyperparameters for each model (except for the MLR and ARIMA models used as baseline). More specifically, the following hyperparameters were tuned:

- Number of layers (up to five were tested)
- Number of units per layer (in the range [32, 512] with step size set to 32)
- Learning rate (sampled uniformly in the range [0.0001, 0.01])
- Number of epochs

For the hyperparameter search, Bayesian optimization was used as tuner. (The Bayesian optimization tuner tries to predict which hyperparameters that are likely to improve the model given previous results [27]). The motivation for this choice is the large number of possible hyperparameter combinations, making it infeasible to test all of them within a reasonable amount of time. Instead, it was assumed that the tuner after 75 trials would find some optimal set of hyperparameters. The hyperparameter search was done in total three times for every model; one search each was performed for data input windows of different sizes, namely 8 h, 16 h, and 24 h. After completing the search, the number of epochs for each model were tuned, and all models were re-trained and evaluated on the validation and test data.

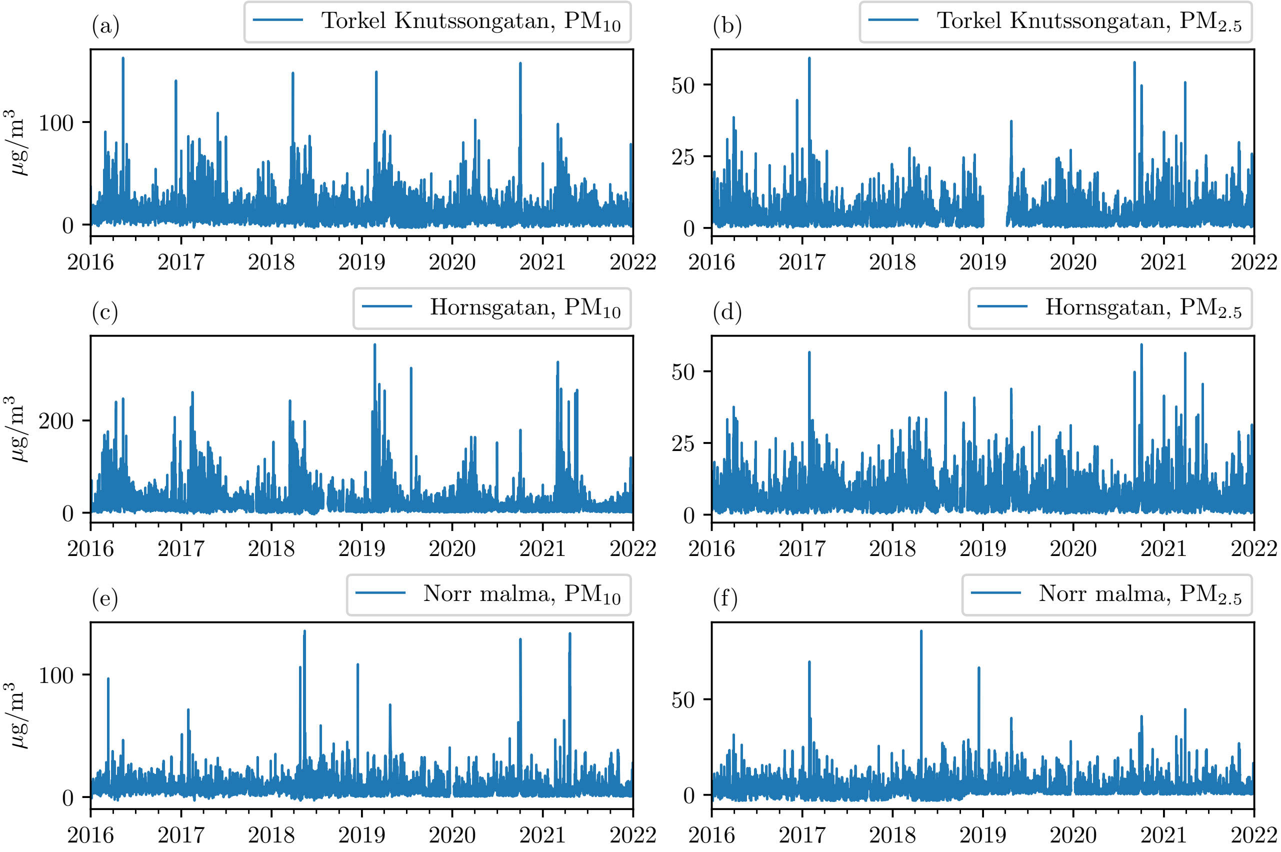

Figure 3.4: Time series plots for PM$_{10}$ and PM$_{2.5}$.

4. Results

4.1 Linear model

Predictions together with actual PM_{10} values for the linear model with different window input sizes (for a two week period) are shown in Figure 4.1. The errors on the validation and tests sets are summarized in Table 4.1. Clearly, the linear model did not benefit from having larger sizes of the input windows as the errors on both the test set and validation set increased. From Figure 4.1, it is also clear that pollution peaks (especially the peak appearing just before 2020-04-03) failed to be predicted.

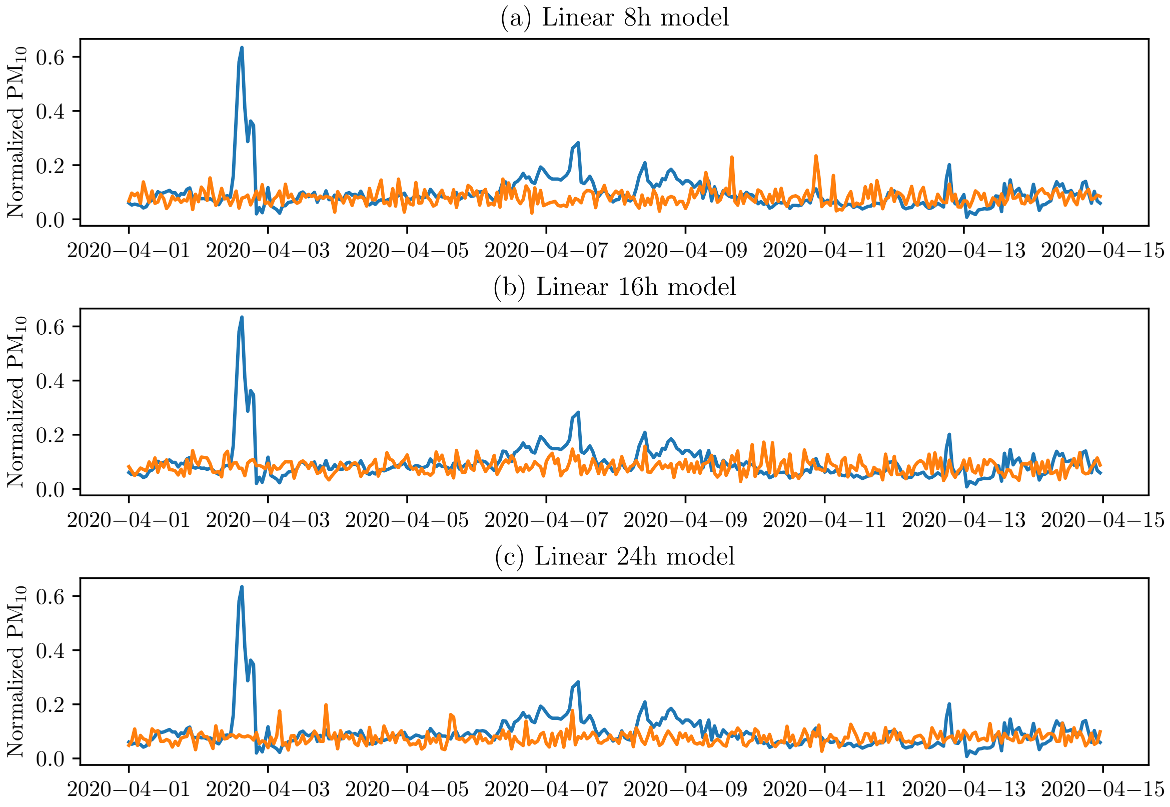

Figure 4.1: Predictions for the linear model during a two-week period.

4.2 Dense model

The error on the training and validation sets for the best models during re-training (see section 3.2) are shown in Figure 4.2. As can be seen, there were no indications

Table 4.1: Errors on validation and test sets for the linear model.

Window input size	Error, validation set	Error, test set
8h	0.0415	0.033
16h	0.0464	0.0373
24h	0.0485	0.037

of over-fitting with either of the input window sizes. Similar as for the linear model, predictions together with actual PM_{10} values (for the same time period) are shown in Figure 4.3, and the errors in the test and validation sets are summarized in Table 4.2. Though the dense models had lower errors than the linear model (e.g. 0.0221 vs 0.0330 for the 8 h input window), the dense model also failed to predict pollution peaks, as is evident from Figure 4.3.

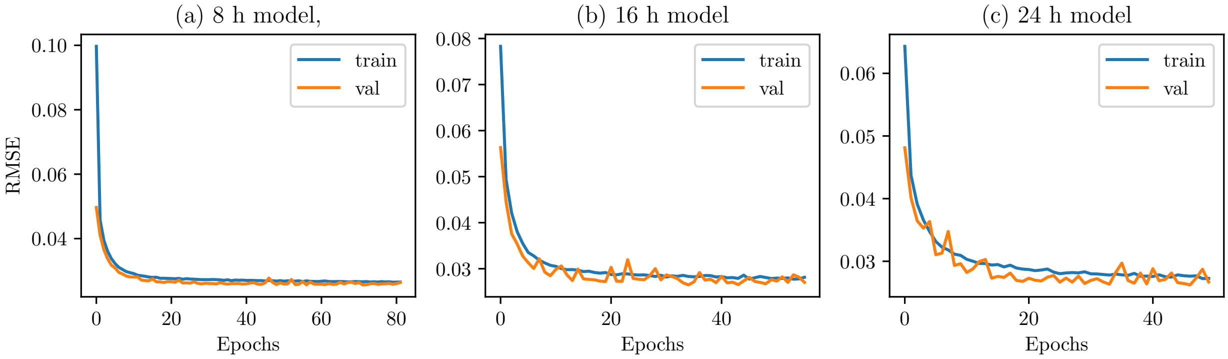

Figure 4.2: Training and validation errors for the dense model.

Table 4.2: Error on validation and test sets for the dense model.

Window input size	Error, validation set	Error, test set
8 h	0.0262	0.0221
16 h	0.0264	0.0218
24 h	0.0272	0.0222

4.3 LSTM model

The validation and test errors for the LSTM model is summarized in Table 4.3.

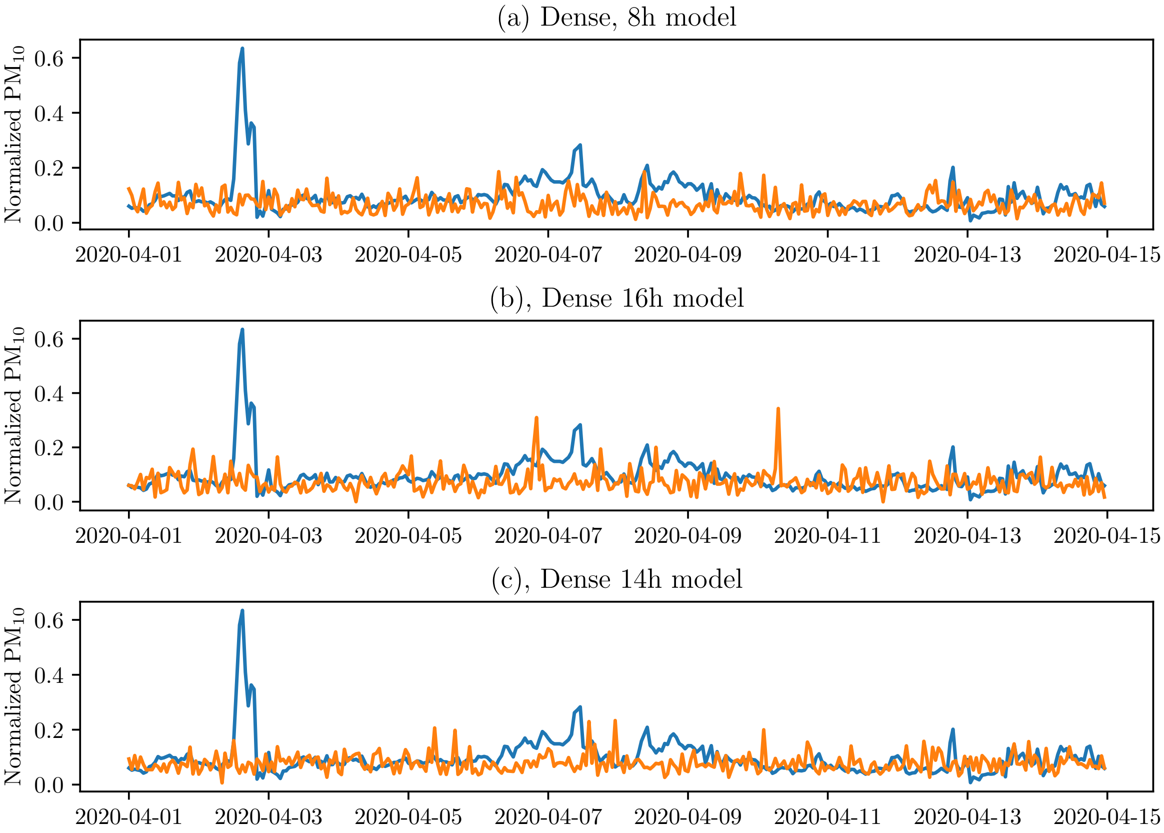

Figure 4.3: Predictions.

Table 4.3: Error on validation and test sets for the LSTM model.

Window input size	Error, validation set	Error, test set
8 h	0.0255	0.0214
16 h	0.0252	0.0211
24 h	0.0254	0.0211

5. Discussion and Conclusions

6. Bibliography

[1] World Health Organization, “Ambient air pollution: a global assessment of exposure and burden of disease,” tech. rep., World Health Organization, 2016.

[2] G. W. VanLoon and S. J. Duffy, *Environmental Chemistry.* London, England: Oxford University Press, 3 ed., Sept. 2010.

[3] SLB-analys, “Luften du andas - nu och de kommande dagarna: Utveckling av ett automatiskt prognossystem för luftföroreningar och pollen,” tech. rep., SLB-analys vid miljöförvaltningen i Stockholm, 2021.

[4] M. El-Harbawi, “Air quality modelling, simulation, and computational methods: a review,” *Environmental Reviews*, vol. 21, pp. 149–179, Sept. 2013.

[5] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, “Deep learning for air quality forecasts: a review,” *Current Pollution Reports*, vol. 6, pp. 399–409, Sept. 2020.

[6] H. Taheri Shahraiyni and S. Sodoudi, “Statistical modeling approaches for PM10 prediction in urban areas; a review of 21st-century studies,” *Atmosphere*, vol. 7, no. 2, 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.

[8] OECD, *The Economic Consequences of Outdoor Air Pollution.* Paris: OECD Publishing, 2016.

[9] R. P. Schwarzenbach, P. M. Gschwend, and D. M. Imboden, *Environmental organic chemistry.* Nashville, TN: John Wiley & Sons, 3 ed., Oct. 2016.

[10] X. Querol, A. Alastuey, C. Ruiz, B. Artiñano, H. Hansson, R. Harrison, E. Buringh, H. ten Brink, M. Lutz, P. Bruckmann, P. Straehl, and J. Schneider, “Speciation and origin of PM10 and PM2.5 in selected european cities,” *Atmospheric Environment*, vol. 38, pp. 6547–6555, Dec. 2004.

[11] “European commission: Air quality standards.” `https://ec.europa.eu/environment/air/quality/standards.htm`. Accessed April 30, 2022.

[12] SLB-analys, “Luften i stockholm, Årsrapport 2021,” Tech. Rep. 2022–5787, SLB-analys vid miljöförvaltningen i Stockholm, 2021.

[13] “Luftövervakning.” `https://www.slb.nu/slbanalys/matningar/`. Accessed April 28, 2022.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer New York, 2009.

[15] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to time series analysis and forecasting.* Wiley Series in Probability and Statistics, Nashville, TN: John Wiley & Sons, 2 ed., Apr. 2015.

[16] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine Learning: A First Course for Engineers and Scientists.* Cambridge University Press, 2022.

[17] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis.* Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Blackwell, 5 ed., Mar. 2012.

[18] J. F. Pucer, G. Pirš, and E. Štrumbelj, "A bayesian approach to forecasting daily air-pollutant levels," *Knowledge and Information Systems*, vol. 57, pp. 635–654, Mar. 2018.

[19] F. Chollet, *Deep learning with python.* New York, NY: Manning Publications, Oct. 2017.

[20] M. Arsov, E. Zdravevski, P. Lameski, R. Corizzo, N. Koteli, S. Gramatikov, K. Mitreski, and V. Trajkovik, "Multi-horizon air pollution forecasting with deep neural networks," *Sensors*, vol. 21, p. 1235, Feb. 2021.

[21] SMHI, "Datavärdskap för luftkvalitet." `https://www.smhi.se/data/miljo/luftmiljodata`. Accessed May 3, 2022.

[22] "Environmental monitoring program area: Air." `https://www.naturvardsverket.se/en/environmental-work/environmental-monitoring/environmental-monitoring-program-areas/air/`. Accessed April 27, 2022.

[23] "Historiska data." `https://www.slb.nu/slbanalys/historiska-data-met/`. Accessed April 27, 2022.

[24] "Time series forecasting." `https://www.tensorflow.org/tutorials/structured_data/time_series`. Accessed April 27, 2022.

[25] A. Gilik, A. S. Ogrenci, and A. Ozmen, "Air quality prediction using CNN+LSTM-based-based hybrid deep learning architecture," *Environmental Science and Pollution Research*, vol. 29, pp. 11920–11938, Sept. 2021.

[26] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting.* Springer texts in statistics, Cham, Switzerland: Springer International Publishing, 3 ed., Aug. 2016.

[27] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, "Keras-tuner." `https://github.com/keras-team/keras-tuner`, 2019.

A. Time series properties of

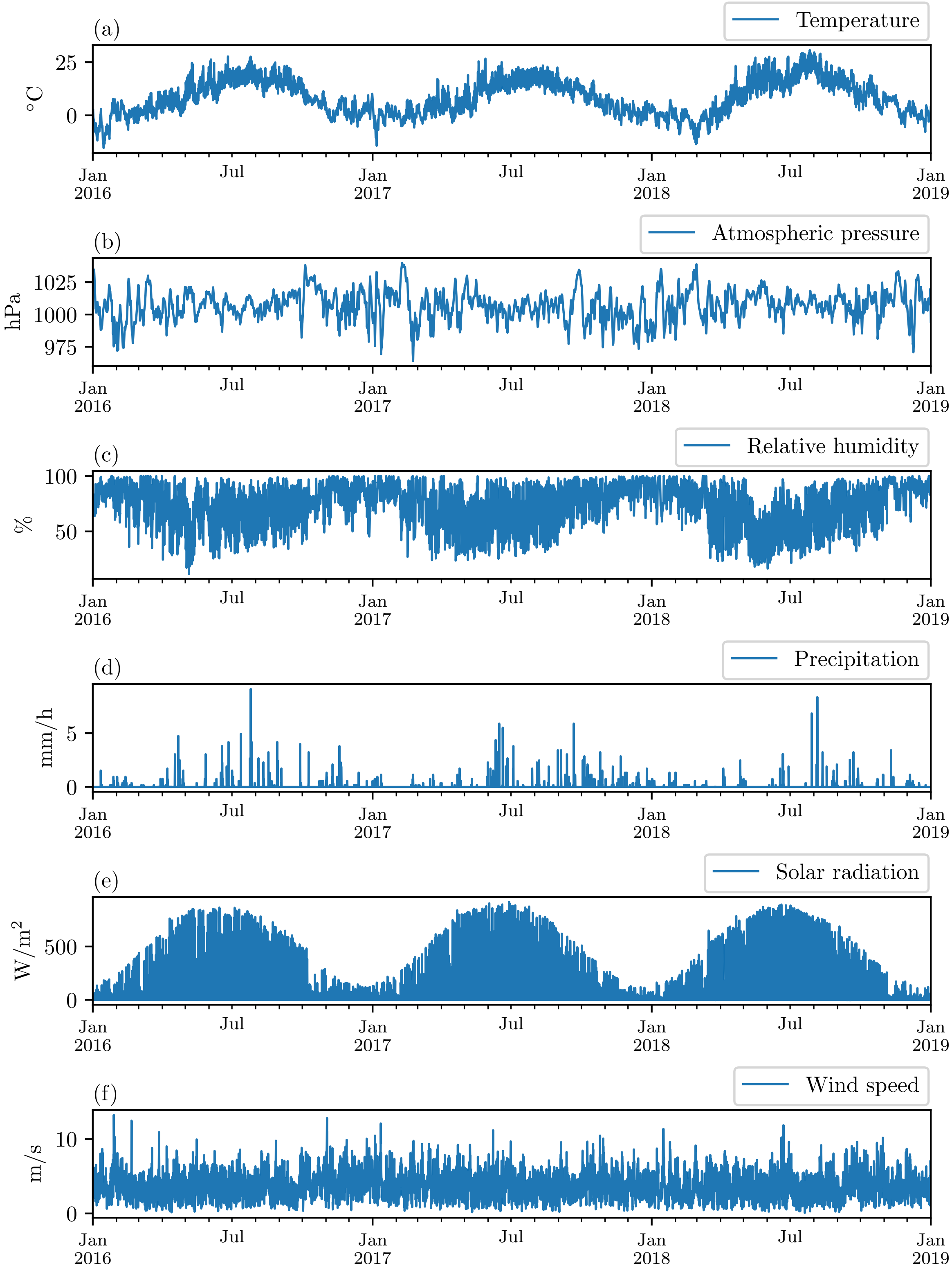

Figure A.1: Time series plots for weather parameters at Torkel Knutssongatan.