https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021

Atmospheric Measurement Techniques
Discussions

Neural network modelling to estimate particle size distribution based on other particle sections and meteorological parameters

Pak Lun Fung[1,2], Martha Arbayani Zaidan[1,3], Ola Surakhi[4], Sasu Tarkoma[5], Tuukka Petäjä[1,3] and Tareq Hussein[1,6,7]

[1]Institute for Atmospheric and Earth System Research / Physics, Faculty of Science, University of Helsinki, Finland; pak.fung@helsinki.fi; martha.zaidan@helsinki.fi; tuukka.petaja@helsinki.fi; tareq.hussein@helsinki.fi
[2] Helsinki Institute of Sustainability Science, Faculty of Science, University of Helsinki, Finland
[3] Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing 210023, China
[4] Department of Computer Science, The University of Jordan, Amman 11942, Jordan; ola.surakhi@gmail.com
[5] Department of Computer Science, Faculty of Science, University of Helsinki, Finland; sasu.tarkoma@helsinki.fi
[6] Department Material Analysis and Indoor Chemistry, Fraunhofer WKI, D-38108 Braunschweig, Germany
[7] Department of Physics, The University of Jordan, Amman 11942, Jordan

Correspondence to: Pak Lun Fung and Tareq Hussein

Abstract.

In air quality research, often only particle mass concentrations as indicators of aerosol particles are considered. However, the mass concentrations do not provide sufficient information to convey the full story of fractionated size distribution, which are able to deposit differently on respiratory system and cause various harm. Aerosol size distribution measurements rely on a variety of techniques to classify the aerosol size and measure the size distribution. From the raw data the ambient size distribution is determined utilising a suite of inversion algorithms. However, the inversion problem is quite often ill-posed and challenging to invert. Due to the instrumental insufficiency and inversion limitations, models for fractionated particle size distribution are of great significance to fill the missing gaps or negative values. The study at hand involves a merged particle size distribution, from a scanning mobility particle sizer (NanoSMPS) and an optical particle sizer (OPS) covering the aerosol size distributions from 0.01 to 0.42 µm (electrical mobility equivalent size) and 0.3 µm to 10 µm (optical equivalent size) and meteorological parameters collected at an urban background region in Amman, Jordan in the period of 1st Aug 2016–31st July 2017. We develop and evaluate feed-forward neural network (FFNN) models to estimate number concentrations at particular size bin with (1) meteorological parameters, (2) number concentration at other size bins, and (3) both of the above as input variables. Two layers with 10–15 neurons are found to be the optimal option. Lower model performance is observed at the lower edge (0.01< Dp< 0.02 µm), the mid-range region (0.15< Dp< 0.5 µm) and the upper edge (6< Dp< 10 µm). For the edges at both ends, the number of neighbouring size bins is limited and the detection efficiency by the corresponding instruments is lower compared to the other size bins. A distinct performance drop over the overlapping mid-range region is due to the deficiency of a merging algorithm. Another plausible reason for the poorer performance for finer particles is that they are more effectively removed from the atmosphere compared to the coarser particles so that the relationships between the input variables and the small particles is more dynamic. An observable overestimation is also found in early morning for ultrafine particles followed by a distinct underestimation before midday. In the winter, due to a possible sensor drift and interference artefacts, the model performance is not as good as the other seasons. The model by meteorological parameters using 5-min data (R^2 = 0.22–0.58) shows poorer results than data with longer time resolution (R^2 = 0.66–0.77). The model by the number concentration at the other size bins can serve as an alternative way to replace negative number in size distribution raw dataset thanks to its high accuracy and reliability (R^2 = 0.97–1). This negative numbers filling method can maintain a symmetric distribution of errors.

Keywords.
Aerosol size distribution, feed-forward neural network, atmospheric aerosols particles, missing data; SMPS; OPS

1 Introduction

Particulate matter (PM) is the principal component of air pollution. PM includes a range of particle sizes, such as coarse (1< Dp<10 µm), fine (0.1< Dp<1 µm), and ultrafine particles (UFP, Dp< 0.1 µm). Through human's inhalation, coarse particles usually are partly deposited in the head airway (5–30 µm) by the inertial impaction mechanism, and are partly deposited in the tracheobronchial region, mainly through sedimentation (1–5 µm). The particles may be further absorbed or removed by mucociliary clearance (Gupta and Xie, 2018). The remaining fine and UFP, due to their high surface area to mass ratios (Kreyling et al., 2004), penetrate deeply into the alveolar region, where removal mechanisms may be insufficient (Gupta and Xie, 2018). Evidence suggests that the adverse associations of short-term UFP exposure with acute and chronic problems ranging from inflammation, exacerbation of asthma, and metal fume fever to fibrosis, chronic inflammatory lung diseases, and carcinogenesis (Spinazzè et al., 2017) might be at least partly independent of other pollutants (Ohlwein et al., 2019). Various studies have demonstrated that inhaled or injected UPF could enter systemic circulation and migrate to different organs and tissues (Xing et al., 2016;Londahl et al., 2014) .

Other than health effects, particles of various sizes also contribute to Earth's ecosystem and climate differently. For instance, fine and UFP are capable of growing up to diameters of 0.02–0.1 µm within a day (Kulmala et al., 2004;Kerminen et al., 2018) where they constitute a fraction of cloud condensation nuclei, and thus, indirectly affecting the climate (Kerminen et al., 2012). The drivers behind aerosol particles vary between natural and anthropogenic as well as primary and secondary. Primary particles are emitted to the atmosphere as particles, such as sea salt or dust particles, while secondary particles form in the atmosphere through gas-to-particle transformation, which has been known as new particle formation (NPF) observed in various environments and contributing to a major fraction of the total particle number budget (Kulmala et al., 2004;Kerminen et al., 2018). In addition, while fine particles cool the climate by predominantly scattering shortwave radiation, coarse particles warm the climate system by absorbing both shortwave and longwave radiation (Kok et al., 2017). Indeed, the complexity of urban aerosols is tribute to the fact that several sources can contribute in the same particle size range (Rönkkö et al., 2017).

Currently, the most commonly reported aerosol variables are particle mass concentration and particle number concentration. The former metric, which is dominated by coarser particles, is included as air quality indicators (e.g. mass concentrations of both thoracic particles PM_{10} and fine particles $PM_{2.5}$); however, it has been argued that this might ignore the potential adverse effect of UFP on health (Zhou et al., 2020). The latter one describes better the distribution of finer particles, but it neglects the influence of coarse particles. Using either particle mass concentration or particle number concentration solely is not enough to fully review the health effects and the Earth's climate system by aerosol particles. Therefore, in order to understand the origin of atmospheric aerosol particles and their potential impacts at a specific location, the whole size distribution of these particles needs to be studied (Zhou et al., 2020).

Recently, due to urbanization and increased population, megacities have increased their contribution to atmospheric aerosol pollution massively Lelieveld et al. (2015). Middle East and North Africa (MENA) regions, with an average annual growth rate of 1.74% in 2019 (World Bank Group, 2019), has one of the world's regions most rapidly expanding populations. With the population of 578 million, several cities in MENA regions are among the 20 most polluted cities in the world. The annual average concentrations of some pollutants, for example $PM_{2.5}$ in MENA (54.0 μg m^{-3}) often exceed 5 times the WHO recommended levels (10.0 μg m^{-3}) (World Health Organisation, 2019). Many countries in MENA are dealing with negative impacts of air pollution in terms of both economic burden and health aspect (Ahmed et al., 2017;Goudarzi et al., 2019). Air Pollution in this region is estimated to cause 133,000 premature deaths annually, almost half of which are attributed to natural sources of air pollution, such as windblown sea salt and desert dust (Gherboudj et al., 2017). Apart from natural pollutants, anthropogenic activities also play a major role in driving the air quality. They include the extensive development of petrochemical industry, vehicular emissions and open burning of waste (Arhami et al., 2018).

However, aerosol studies in this region have not paid attention to the aerosol number size distribution so far. Among the few studies published, most report mass concentration (Goudarzi et al., 2019;Arhami et al., 2018;Borgie et al., 2016), while some focused on the total particle number in MENA regions. Studies on the size-fractionated number concentrations are, nonetheless, scarce (e.g. Hakala et al., 2019) due to the unavailability of instruments for measuring UFP in many air quality monitoring stations (Spinazzè et al., 2017). Determining aerosol number size distribution for a wide size range in a reliable manner is a challenging task. The fact that the ambient distributions range from nanometers to several micrometers dictates the use of multiple sizing techniques. For the sub-micron size range, electrical mobility equivalent diameter is commonly used as the size parameter and the measurements are performed with Differential Mobility Particle Sizer (DMPS) or Scanning Mobility Particle Sizer (SMPS) instruments (e.g. Wiedensohler et al., 2012). These systems determine the aerosol size according to electrical mobility equivalent size. The larger particles (approximately > 0.3 μm) can be classified according to their aerodynamic or optical size (Kulkarni et al., 2011). In order to obtain the full aerosol size distribution, this data needs to be merged. Unfortunately this task is not trivial as the merging requires knowledge on the chemical composition (influencing the refractive index and thus the optical size), shape (influencing electrical mobility equivalent size), or effective density (influencing aerodynamic size) (Kannosto et al., 2008).

In addition, the raw data from these instruments must be inverted to obtain the particle size distribution. This is not a straightforward problem. A proper inversion algorithm is required to restore the particle size distribution using the recorded kernel function between the raw response and the size distribution (Cai et al., 2018). Depending on the instruments used and the measurement environments, some use a built-in inversion algorithm in the instruments. Some develop their own inversion methods; however, they all have their drawbacks. Examples include that the least square method may magnify the random errors in CPC raw counts into relatively large uncertainties (Enting and Newsam, 1990), the stepwise method may cause non-negligible errors (Lehtipalo et al., 2014), and that the smoothing step method may introduce bias in the shape of the inverted distribution function. (Markowski, 1987). Kandlikar and Ramachandran (1999) pointed out that there is not a single universal inversion algorithm applicable to all situations.

In this study, the built-in inversion algorithm was used. Especially in the size range of low number concentration, this algorithm can lead to negative values when the kernel functions are not optimally configured. These negative values have no physical meanings. Some might just omit the negative values or simply use nearest neighbour linear interpolation to replace the negative values. However, the former method might cause asymmetric error for very small measured number concentration values (Viskari et al., 2012), while the latter could result in too high values concurrently. To fill this knowledge gap, statistical models can serve as an alternative to estimate of size-fractioned number concentration by using other available measurements.

Similar to other air quality parameters, modelling of size-fractionated particle number concentrations have been increasingly brought into the spotlight because of its potential health hazards. One of the most commonly used data-driven methods, generalised linear regression models, have been extensively utilised in modelling size-fractionated particle number concentrations, for example, in urban regions in Helsinki, Finland (Clifford et al., 2011;Hussein et al., 2007), in Toronto, Canada (Sabaliauskas et al., 2012), in Brisbane, Australia (Rahman et al., 2017), and in three cities in Germany (Gerling et al., 2020). Besides linear regression models Mølgaard et al. (2013), refined the statistical model using Bayesian inference and autoregressive parameters in five European cities Reggente et al. (2014) improvised by Gaussian process based on the measurements of oxides of nitrogen in Antwerp, Belgium. These approaches are categorised as transparent machine learning (ML) processes, as known as white-box (WB) models, from which one can clearly explain how they behave, how they produce predictions and what the influencing variables are (Rudin, 2019). Another data-driven approach black-box (BB) models, which refer to ML systems being viewed in terms of its inputs and outputs, without any knowledge of its internal workings or underlying principles (Rudin, 2019). They are considered to work generally better in terms of accuracy, but provide limited transparency and accountability regarding the results (Zaidan et al., 2019;Fung et al., 2020). One example of BB model is artificial neural network (ANN), which were applied extensively to estimate other air pollutant parameters (Freeman et al., 2018;Cabaneros et al., 2019). They provide a robust approach for approximating complex functions due to its ability to mimic non-linearity of the functions and its well-developed optimisation. Al-Dabbous et al. (2017) has demonstrated the use of ANN to estimate three ranges of UPF at a roadside site in Fahaheel, Kuwait. They addressed the importance of including meteorological parameters in the modelling process, which was later validated by Zaidan et al. (2020) who estimated daily and hourly total particle number concentration by only a combination of meteorological parameters in Amman, Jordan.

The objectives of the paper is estimate aerosol total number concentration from meteorological observations and to advance the previous study by Zaidan et al. (2020) with a finer temporal and size-bin resolution. In order to do so, we place emphasis on to estimate particle number concentration of a specific size bin by the interaction with other size bins and meteorological variables. In this study, we propose three approaches in terms of different input variables when we carry out the modelling: (1) only meteorological parameters, (2) only particle size distribution, and (3) both particle size distribution and meteorological parameters. Based on the general data analysis of the particle size distribution and the meteorological condition, we further explain the source of different size bins at certain weather conditions and the correlation among the particle size distribution and meteorological parameters in Section 3. Evaluation of models is discussed in Section 4, in terms of its diurnal cycle, weekend effect and seasonal variation. We also examine the possible technical reasons for the pattern found and the application of the models.

2 Methods

2.1 Measurement sites and Instruments

In this study, we collected a dataset obtained from a measurement campaign in Amman, the capital city of Jordan, between 1 August 2016 and 31 July 2017. The city represents an area with Middle Eastern urban conditions within the Middle East and North Africa (MENA) region. This region serves as a compilation of different aerosol particle sources including natural dust, anthropogenic pollution (e.g., generated from the petrochemical industry and urbanization), as well as new particle formation.

The database includes particle size distribution and meteorological parameters, as mentioned in the first step in Figure 1. The aerosol measurement was carried out at the aerosol laboratory located on the third floor of the Department of Physics, University of Jordan (32°00′ N, 35°52′ E) in the neighbourhood of Al Jubeiha. The campus is situated at an urban background region in northern Amman. In particular, the campaign measured the particle number size distribution using a scanning mobility particle sizer (NanoScan SMPS 3910, TSI, MN, USA). It monitors the particle size distributions as electrical equivalent diameter 0.01–0.42 μm (13 channels). The size range of the SMPS system can be extended to coarse particles with an additional compact instrument: an optical particle sizer (OPS 3330, TSI, MN, USA). OPS measures optical diameter 0.3–10 μm (13 channels). This optical sizing method reports an optical particle diameter, which is often different from the electrical mobility diameter measured by the SMPS technique. The measurements were combined to provide a particle size distribution of wider particle diameter range 0.01–10 μm, which is further described in Section 2.2. The SMPS inlet flow rate was 0.75 lpm (±20%) while the sample flow rate was 0.25 lpm (±10%). The flow rate of OPS was about 1 lpm. The aerosol transport efficiency through the aerosol inlet assembly was determined experimentally: ambient aerosol sampling alternatively with and without sampling inlet, and the aerosol data was corrected accordingly. The penetration efficiency was ~47% for 0.01 μm, ~93% for 0.3 μm and ~40% for 10 μm (Hussein et al., 2020). These deficiency of measurement at the upper and lower edges is somewhat in alignment with other literatures. Particle size measured by nanoSMPS (Tritscher et al., 2013) tended to be underestimated for spherical particles larger than 0.2 μm by up to 34% (Fonseca et al., 2016). Liu et al. (2014) clearly portrayed that the detection limit of particle size below 0.03 μm is about 80–500 cm^{-3}, which is up to 10 times larger than that of coarser particles, for other versions of SMPS. Stolzenburg and McMurry (2018) explained that discrepancies could be resulted from DMAs with transfer functions that were degraded (i.e., broadened) by flow distortions caused by particle deposition within the classifier tube, sizing errors due to errors in flowmeter calibrations or leaks, CPC concentration errors due to improper pulse counting, and continuity failure in the DMA high voltage connection.

The meteorological measurement was performed with a weather station (WH-1080, Clas Ohlson: Art.no.36-3242, Helsinki, Finland) with a time resolution of 5 minutes. The meteorological data were comprised of ambient temperature (Temp, resolution 0.1°C), relative humidity (RH, resolution 1%), wind speed (WS), wind direction (WD, 16 equal divisions) and air pressure (P, resolution 0.3 hPa) (Hussein et al., 2019;Hussein et al., 2020;Zaidan et al., 2020). Wind direction is resolved into north and east direction, as WD-N and WD-E, respectively. The data collection process is illustrated in the first step in the database block in Figure 1.

2.2 Data pre-processing

The next step in Figure 1 is data pre-processing. Since the sampling time resolution of SMPS and OPS was 1 min and 5 min, respectively, we synchronised the data into 5-min average. Since a part of the size ranges in both instruments are overlapping with each other, the last two size bins in SMPS and the first size bin in OPS were neglected. Finally, we merged the size range of electrical mobility diameter 0.01–0.25 µm by SMPS and optical diameter 0.32–10 µm by OPS, and obtain a wider particle size distribution which covers the diameter range 0.01–10 µm. Merging electrical mobility diameter and optical diameter can be a challenge and the overlapping region is often calculated with high uncertainty (DeCarlo et al., 2004;Tritscher et al., 2015). The challenge arises because the optical diameters are measured based on the refractive index of the particles, which depends on their chemical composition. Therefore the sizing will vary over time. There is also a very slight dependency with the SMPS system that is linked to the shape of the particles, which influences their sizing.

We also calculated the particle number concentration with four particle diameter modes (size-fractionated number concentration): nucleation (0.01–0.025 µm), Aitken (0.025–0.1 µm), accumulation (0.1–1 µm) and coarse mode (1–10 µm). Subsequently, the total number concentration was obtained as the sum of all these fractions. The size-fractionated number concentrations were obtained by summing up the measured particle number size distribution over the specified particle diameter range.

In order to perform neural network modelling, aerosol and meteorological data were first linearly interpolated in case of short missing data periods. For missing data over longer periods, the whole rows are eliminated. The shorter missing data occurs due to technical faults while the longer missing periods are attributed to instrument maintenance (Zaidan et al., 2020). Only 71.8% of total data was retained for modelling in the measurement period. Since the data were obtained from different measured variables with various physical units and magnitudes, it was crucial to normalise the data. The scaling factor depends on which activation function is chosen. In this case, the datasets were scaled so that it has a mean of 0 and a standard deviation of 1 to transform them into the range of the activation function. The standardised data was then separated into different months for the reason of the seasonal variation in the atmospheric condition. The data was further divided into training set (70%) and testing set (30%). The processed data were also converted to hourly and daily averages for reporting purposes.

2.3 Modelling

After data collection and data pre-processing procedures, the next step is model optimisation (Figure 1). ANN models have been utilised in predicting air quality (Freeman et al., 2018;Maleki et al., 2019;Cabaneros et al., 2019;Zaidan et al., 2020). Neural networks provide a robust approach for approximating real-valued target functions because they can mimic the non-linearity of the functions and their optimization methods are well developed (Zaidan et al., 2017). The architecture of neural networks consists of nodes which generate a signal or remain silent as activation function (Figure 2). Activation function in each layer determines the output value of each neuron that becomes the input values for neurons in the next hidden layer connected to it. In this paper, feedforward neural network (FFNN) is used instead of a more sophisticated time delay neural network (TDNN) because some of the rows in the dataset were removed in the data pre-processing step due to the existence of missing data and TDNN cannot be performed without time continuity. FFNN usually consists of

https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021

a series of layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the network's output. A neuron can be thought as a combination of two parts:

$$z_j^{(L)} = \sigma(\sum_{i=1}^{n} w_{ji}^{(L)} x_i + b_j^{(L)}) \quad (1),$$

where $z_j^{(L)}$ and $b_j^{(L)}$ are the intermediate output and the bias term for the jth neuron at Lth layer, respectively. $w_{ji}^{(L)}$ is the jth weight for each data points x_i at Lth layer. The second part performs the activation function (sigmoid function in this study) on z_j to give out the output of the neuron:

$$\sigma\left(z_j^{(L)}\right) = \frac{1}{1 + \exp^{-z_j^{(L)}}} \quad (2),$$

The FFNN model was created, trained and simulated with MATLAB (version: 8.3.0.532), using Neural Network Toolbox. We initialised the weights randomly and the weights are updated through ''Levenberg-Marquardt'' algorithm optimisation that was the fastest available back-propagation training function (Chaloulakou et al., 2003). We performed several iterations within a cycle to minimise the training loss with Bayesian regularisation. These steps were done iteratively until the best combination of the number of hidden layers and the corresponding number of neurons that provided the minimum error was found. According to the review paper by Cabaneros et al. (2019), a shallow neural network with one hidden layer and enough neurons in the hidden layers can fit any finite input-output mapping problem for non-linear relationship. In the network training process, the number of neurons varied from 2 to 10 neurons per layer with an incremental factor of 2 neurons in each simulation, and from 10 to 25 per layer with an incremental factor of 5 neurons in each simulation. To keep the model simple, we consider only one or two layers in the simulation process because the computing requirements could rise exponentially with the number of layers and neurons. Once we pick the suitable model configuration, the model estimates number concentration using testing data. Finally, the selected performance metrics, described in Section 2.4, can be calculated and we evaluate which approach is the most suitable for size distribution estimation.

2.4 Performance metrics

We choose the optimal combination of the number of hidden layers and the corresponding number of neurons by checking its mean absolute error (MAE), which is a simple way to illustrate the residuals of the estimated values by the model. In order to identify which size bin manage to be predicted best, two metrics are used, namely coefficient of determination (R^2) and normalised root-mean-square error (NRMSE). R^2 measures how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. NRMSE represents the standard deviation of the estimated errors with respect to its mean. NRMSE is used rather than commonly used RMSE because the number concentrations of the different size range are of different magnitudes. The comparison in different size range becomes different if RMSE is not normalised with its mean.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (4)$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}}{\bar{y}} \quad (5)$$

where y_i, $\hat{y}_i$ and $\bar{y}$ represent the ith measurement value, the yth estimated value by the model and the mean of the all the measurement data, respectively. n notates the total number of the valid measurement data.

3 General data analysis

3.1 Environmental condition

Hussein et al. (2019) and Zaidan et al. (2020) investigated and described the effect of local weather conditions, respectively. Here we describe briefly the meteorological conditions during the measurement period as background information. Starting from August 2016, the daily temperature decreased gradually from 40°C to its tough 0°C in February 2017. It rose gradually to 40°C in August 2017. During the measurement period, the hourly median value was 19.9°C (Figure 3a). RH varied quite a lot from 10% to 100%, with an hourly median of 52.3%, and did not seem to have a seasonal pattern (Figure 3b). In summer months, wind appeared be stronger but the wind direction is more stable, mostly from northwest (270º–360º). In cold months, averaged wind speed was lower but wind blew from fluctuating direction. During the whole measurement period, wind speed ranged between 0–6 m s^{-1} and its median is 1.39 m s^{-1} (Figure 3c,d). Air pressure varied in a range from 892 to 912 hPa and its hourly median was 900 hPa In spite of the narrow range of variation, winter months seem to have slightly higher air pressure than summer months (Figure 3e).

Meteorological conditions have been suggested to influence particle number concentration. Hussein et al. (2019) demonstrated that number concentration had a rather complex relationship with temperature. Furthermore, number concentration of submicron had a decreasing trend with respect to the wind speed which indicates that most of the submicron fraction is originated from local sources such as combustion processes. Meanwhile, the number concentration of coarse particles had higher concentrations at stagnant conditions and when the wind speed is higher than 5.5 m s^{-1}. It is mainly because of road dust resuspension and might also be attributed to dust storm via long-range transport (Hussein et al., 2019). In this study, we further explore how wind direction influences the particle number concentration (Figure 4). Wind coming from the northwest (225º–325º) was generally stronger, but lower particle number concentration was detected because the measurement area is at the outskirt of downtown. Wind from East and South (45º–225º) has a lower wind speed but a more intense hourly particle number concentration can be detected. From that direction situates the urban city where all kinds of industrial activities take place. When considering only coarse particles, relatively high number concentration is found when south-westerly wind is strong. This can further serve as an evidence that the source of coarse particles in that region might come mostly from long range sea salt from Dead Sea or dust particles from nearby deserts.

3.2 General pattern of particle size distribution

Hourly total number concentration ranged from 1.90×10^3 cm^{-3} to 1.52×10^5 cm^{-3}and its median was 1.36×10^4 cm^{-3}. Figure 5a performed moderate seasonal pattern in general: lower in summer months and higher in colder months. Hussein et al. (2019) also characterised the modal structure of the particle number size distribution for the same site. Four modes have been detected by lognormal fitting, as known as DO-FIT algorithm and modal structure (Hussein et al., 2005;Hussein et al., 2019), revealed that the mode number concentrations of the nucleation, Aitken, and coarse modes were lognormally distributed around their geometric mean values: 0.022 µm, 0.062 µm, and 2.3 µm respectively. However, the accumulation mode number concentration had two distinguished modes with particle diameter centred at 0.017 µm and 0.39 µm. As seen in Table 1, the total number concentration of all particle size (1.70±1.26×10^4 cm^{-3}) is mostly accounted by Aitken mode (45–80%, average: 1.09±1.01×10^4 cm^{-3}), followed by nucleation mode (10–50%, average: 0.48±0.32×10^4 cm^{-3}). Accumulation mode (0–15%, average: 0.13±0.08 cm^{-3}) comes third and only less than 0.5% of the total particle number concentration contain coarse particles with an average of 2.13±2.80 cm^{-3} (Figure 5b-e). Seasonal pattern of the total number concentration resembles the Aitken composition: lower proportion in summer months and higher in colder months. The ratio of nucleation mode performs in an opposite way. The seasonal variation of total number concentration is due to the more suppressed boundary layer in winter (Teinilä et al., 2019) and the elevated wood combustion (Hellén et al., 2017). The particle number of accumulation and coarse mode steadily stay at a low proportion line, which did not account for the total number concentration. It is also noticed that dust episodes occurred with the concentrations that often exceeded 2 cm^{-3} and the daily concentration in the course of these episodes can rise to 20 cm^{-3}. These episodes were often found in spring from February to May and some episodes can last for up to one week.

Similar to many other urban environments, the diurnal pattern observed in this study reflects the combustion emissions from traffic activity, which is more during the workdays (Hussein et al., 2019). The two peaks of the nucleation mode and Aitken mode in the cold months are relevant for the morning and the afternoon traffic rush hours, which are similar to those noticed in most cities in other countries. In warmer months, the diurnal cycles are not as distinct, but a sharp peak of nucleation mode around noon is found, which is associated with the occurrence of new particle formation. These events occurred very often in the summer as suggested by Hussein et al. (2020). The amplitude of diurnal cycles of coarse mode is small while the patterns of accumulation are not clear (Figure 6).

3.3 Correlation analysis

Figure 7 demonstrated the interaction among the whole measured spectrum shows three range clusters based on their correlation with the number concentration at other bin sizes: 0.01–0.205 μm, 0.205–0.875 μm and 0.875–10 μm. 0.01–0.205 μm and 0.875–10 μm fall entirely within the size range detected by SMPS and OPS, respectively. The 5-min number concentration of smaller size and bigger size bins have clear and strong correlation with the concentration of its neighbouring size bin. However, particles of size 0.205–0.875 μm are located in the overlapping regions by the two instruments; as a result, do not correlate well with other size bins. The correlation of 5-min particle size distribution with meteorological parameters are generally low. Temperature appears to be the most correlated parameters for all bin sizes among all the parameters we used in this study. Smaller size range have higher Pearson's correlation coefficient (R) than larger size range for WD, WS and P.

The 5-min averaged data show similar correlation for the particle size distribution except for the smallest size bin. The hourly and daily data have higher correlation with the other size bins which are also monitored by SMPS. The 5-min averaged data show different correlation from the hourly and daily averaged data performed by (Zaidan et al., 2020). The correlations of 5-min size distribution with all meteorological variables are below 0.5 for all size range. However, for hourly and daily averaged data, R is much higher in specific size bins. Hourly and daily temperature, in particular, show increasing R with larger particle size for accumulation and coarse mode. Overall, the correlations increase with the longer averaging windows. This might be due to the buffer period the meteorological conditions act on the dispersion of particles. Based on this result, using data with finer temporal resolution might be considered to be less influential to the accuracy of modelling.

4 Model Evaluation

4.1 General evaluation

Figure 8 illustrates how well the models of the three approaches perform in term of R^2 and NRMSE.

Approach 1 (Size distribution prediction based on meteorological parameters only): For more than half out of the 23 size bins, 2 layers and 15 neurons is the best combination where the residuals are the lowest (Table 2). Since the poor correlation with meteorological condition, we expect a low correlation of determination even using the optimal configuration neural network (R^2 = 0.22–0.58). The R^2 looks poor at the nucleation mode (0.01< Dp <0.03 μm) of the whole size distribution around nucleation mode (R^2~0.2). The rest of the size bins have better and stable performance (R^2 = 0.4–0.58). This shows that the instrument might have a poor detection efficiency for particles of smaller size. By using FFNN, the model performance of 5-min data for all size bins (R^2 = 0.22–0.58) is worse than using daily data (R^2 = 0.77) performed in Zaidan et al. (2020). Compared with hourly data (R^2 = 0.66), the overall model performance of 5-min data is comparable (R^2 = 0.67).

Approach 2 (Mutual size distribution prediction based on other particle sections only): work well with most combination of number of layers and neurons. They did not show a clear difference among the combinations we choose. There is no single combination which entirely outperform the others in all size bins. We summed up the MAE for all size bins and decided to stick to 2 layers and 10 neurons with the overall lowest residuals (Table 2). R^2 are all above 0.97 for all bin sizes, and NRMSE is 0.01–0.25 for bin sizes. The results are expected because there are 22 inputs and one output. Relatively worse correlation at the edges of size bins (0.01< Dp< 0.02 μm; 6< Dp< 10 μm) is found because of the lack of nearby size bins which has high correlation with the corresponding size bin. Another reason could be that the instrument has a higher detection limits for smaller particles (Liu et al., 2014). The poorer performance for smaller size might be due to a coarser sizer resolution compared to other SMPS components (Tritscher et al., 2013), so that NanoSMPS does not reflect the real enough size distribution in the atmosphere. Relatively poor modelling performance at the middle size range (0.15< Dp< 0.5 μm) in the whole measured spectrum is because of the overlapping of instruments. This also ascertain the importance of creating a better algorithm when we merge two or more size distribution by different instruments. In this study, the measuring techniques and the measuring targets are different by the SMPS and OPS. The merging of the two measuring targets, the optical particle diameter and the electrical mobility diameter, might create significant uncertainties (DeCarlo et al., 2004;Tritscher et al., 2015). The estimation of certain bin size by other bin sizes can be thought of replacing negative values in the raw data by particle sizers. While some instrument manufactures create built-in algorithms to replace with artificial non-negative numbers, most end-users simply remove the seemingly impossible negative values from the dataset. The perfect way to do it is to have a parallel instrument that overlaps with that particle size range. However, in many cases, this is not possible as a result of financial constraints. Therefore, we shall rely on the mutual relationship between the size sections in the aerosol population. Negative values appear often at size bins with very low number concentration (usually in coarse mode). Instead of eliminating them, this alternative could maintain the symmetry of the error distribution of the number concentration (Viskari et al., 2012) and minimise the uncertainties caused.

Approach 3 (Mutual size distribution prediction based on meteorological parameters and other particle sections): the general results are similar as in PSD. However, the more input variables do not enable the model to work better. At some bin size the R^2 are even slightly smaller than PSD solely. Since meteorological data show low correlation with most portion of measured spectrum. In that approach, the addition of meteorological parameters is not beneficial to the modelling process. Due to the lack of improvement in the model development, we will only focus on the two models: met and PSD from now on.

From the perspective of physics, particles in the nucleation mode (0.01< Dp< 0.03 μm) are more sensitive to transformation processes due to their volatility and rather unstable nature (Morawska et al., 2008). This leads to a relatively short lifetime in the atmosphere (Al-Dabbous et al., 2017), therefore the relationships between the input variables and the nucleation mode are not well established. Al-Dabbous et al. (2017) demonstrated that accumulation mode particles (0.1< Dp< 0.3 μm) have much longer lifetimes compared to smaller particles, causing them to be transported for larger distances (Laakso et al., 2003); therefore, the mapping of the relationships between long–range transported accumulation mode particles and covariates is supposed not to well understood. However, the relative prediction ability in this study is not lower given that local meteorological variables were used as input variables. The possible reason is that this mode falls exactly in the instrumental overlapping regions, which leads to a lower predictively. The locally-produced Aitken mode particles (0.03< Dp< 0.1 μm) are less effectively removed by transformation processes (e.g., evaporation and coagulation) from the atmosphere, compared with nucleation mode (0.01< Dp< 0.03 μm), allowing the prediction models to better understand their relationships with the input variables, which is in alignment with Al-Dabbous et al. (2017).

4.2 Temporal pattern

Figure 9 shows the diurnal discrepancies during workdays and weekends. Relative particle number concentration was defined by the modelled concentration with respect to the measured concentration. Values above 1 indicates overestimation while values below 1 suggests underestimation. For approach 1, except for the overlapping size bin, which are underestimated by more than 50% at all time range, the difference between modelled and measured hourly number concentration is within 50% during both workdays and weekends. Overestimation is found in early morning before 3 a.m. during workdays for all size bins, especially for UFP. Following the overestimation, at about 6 a.m. in the morning, the modelled number concentration appears to understate by up to 40%, especially at size bins below 0.1 um. Along the day, the modelling uncertainties are rather small until in the evening from 6 p.m. to 11 p.m. where modelled UFP number concentration show moderate overestimation one more time. It reveals that the model with only meteorological parameters as inputs fail to catch the diurnal pattern from 6 p.m. to 7 a.m. in particular for UFP. The pattern of the performance for weekends does not appear to be as distinctive as on workdays. It shows the overestimation not only for UFP in early morning about 3 a.m., but also at the upper edge larger than 5 um from 3 a.m. to 4 p.m.. At 7.p.m. onwards until noon, an underestimation is found at all size bins. For approach 2, except the overlapping size bin, which has a significant overestimation from 6 p.m. to 7 a.m., most show trivial 10% uncertainty during both workdays and weekends. The model performance over weekends show relatively stronger uncertainties. The smallest bin at 0.01 μm is slightly understated for all hours of a day. Other than these, models with the full spectrum of size distribution as inputs manage to catch fairly well the diurnal pattern for all size bins.

Figure 10 further shows the monthly deviation in modelling performance. For approach 1, higher R^2 is found in November, February and April in the range of SMPS. Other than that, no observable variation in R^2 in approach 1. For approach 2, except in January when all the rows were eliminated because of the lack of wind information, performance in the other months is steady for most size range. At 0.21 μm, the difference in model performance varies across different months. R^2 in winter months are 0.76, 0.36 and 0.61, in November, December and February, respectively, while R^2 exceeds 0.9 in other months. This unexpectedly low R^2 only occurs in the winter months at the overlapping size range. It can be speculated that the measurements by the two instruments differ in a larger extent during winter. This might be attributed to sensor drift and a number of interference artefacts for particle measurements associated with several factors, such as relative humidity, temperature and other gas-phase species, which were demonstrated by several researchers (Lewis et al., 2016;Popoola et al., 2016). Another reason for the difference in model performance can be that the percentage of complete rows in these months are lower than the other months. The drop in data points might impose an influence to the modelling performance. Especially in June, at the few size bins close to the larger edge, R^2 ranges from 0.9 to 0.7. Besides that, some low R^2 can be also found in individual month at both edges of size range, which does not appear to show any patterns.

In short, the prediction ability for lower edge (0.01< Dp <0.03 μm) is found worse in both models. The model performance in mid-range (0.15< Dp< 0.5 μm) and upper edge (6< Dp< 10 μm) are relatively worse for model with other fractionated size bins as input variables according to the aforementioned statistical performance indicators. All statistical prediction simulations are based on the previous history of relationships between the inputs and outputs. As a result, the prediction simulations for different size ranges have significantly unique connections. The model by meteorological parameters considers only 6 predictor variables so the accuracy is lower than the model by PSD. It might not seem surprising that the deviations between the measured and estimated size distribution were not substantial (R2> 0.97, NRMSE< 0.25) because the PSD model take 22 other size bins as predictor variables. This, still, gives a clue that the proposed model can provide adequate solutions to particle size distribution prognostic demands. The estimation of certain bin size by other bin sizes can be thought of replacing 'negative' values in the raw data by particle sizers, including SMPS we used in this paper. Instead of eliminating the negative values, they can be estimated by other size bins with a high accuracy in order to keep the symmetry in data error distribution (Viskari et al., 2012).

5 Conclusion

This paper presents the evaluation of feed-forward neural network (FFNN) models for estimating particle number concentration at various particulate size bins. Input predictors include a merged particle size distribution, by a scanning mobility particle sizer (NanoSMPS) and an optical particle sizer (OPS), which covers size range from 0.01 to 10, and meteorological parameters, including temperature (Temp), relatively humidity (RH), wind speed (WS), wind direction (WD) and ambient pressure (P). The measurements were collected in an urban background region in Amman, the capital of Jordan in the period of 1st Aug 2016–31st July 2017. The total number concentration ($1.70 \pm 1.26 \times 10^4$ cm^{-3}) in the measurement period show moderate seasonal variability owing to the more suppressed boundary layer (Teinilä et al., 2019) and the elevated wood combustion (Hellén et al., 2017) in wintertime. Similar to many other urban environments, the diurnal pattern observed in this study reflects the traffic activity, which has a more pronounced pattern during workdays (Hussein et al., 2019). The amount of coarse particles is trivial in terms of number concentration but dust episodes were found often in spring during the measurement period.

We proposed three approaches with different input variables: (1) only meteorological parameters, (2) only number concentration at the remaining size bins, and (3) both of the above. We performed optimisation to obtain the optimal configuration of the FFNN models, which are two layers with 10–15 neurons, balancing the accuracy and the computing resources. The 5-min averaged meteorological parameters give varying number concentration estimation for various size bins (R^2 = 0.22–0.58), which is outperformed by hourly and daily averaged data (R^2 = 0.66–0.77), as demonstrated by

(Zaidan et al., 2020). The models using the number concentration at the remaining size bins, both with or without meteorological data, show expected perfect performance ($R^2 > 0.97$).

Relatively poor model performance is found in three regions. At the lower edge (0.01< Dp< 0.02 µm) and the upper edge (6< Dp< 10 µm), the number of neighbouring size bins is limited and also the detection efficiency by the corresponding instruments is lower compared to the other size bins. Another noticeable region (0.15< Dp< 0.5 µm) is the overlapping section measured by the two particle sizers and the reason is because of the deficiency of merging algorithm. For all the above approaches, the poorer performance for smaller particles in the nucleation mode could be due to the fact that it is more effectively removed from the atmosphere compared to other modes (Al-Dabbous et al., 2017). An observable overestimation is also found in early morning for ultrafine particles followed by a distinct underestimation before midday. A larger derivation between the measured and the estimated number concentration is found in the winter, which might be caused by sensor drift and interference artefacts (Lewis et al., 2016;Popoola et al., 2016). Despite the high number of input predictors, the good model performance provides an alternative method to fill up the negative values in size distribution raw dataset, which often exist due to ill-configured problems. Instead of removing the factually impossible data point, this way of replacing negative numbers can maintain a symmetric distribution of errors (Viskari et al., 2012) and minimise the uncertainties caused.

Code/Data availability

The code and data is available upon request.

Author contribution

TH and MZ designed the experiments and TH carried them out. PLF and OS developed the model code. PLF prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

References

Ahmed, R., Robinson, R., and Mortimer, K.: The epidemiology of noncommunicable respiratory disease in sub-Saharan Africa, the Middle East, and North Africa, Malawi Med J, 29, 203-211, 10.4314/mmj.v29i2.24, 2017.

Al-Dabbous, A. N., Kumar, P., and Khan, A. R.: Prediction of airborne nanoparticles at roadside location using a feed-forward artificial neural network, Atmos Pollut Res, 8, 446-454, 10.1016/j.apr.2016.11.004, 2017.

Arhami, M., Shahne, M. Z., Hosseini, V., Haghighat, N. R., Lai, A. M., and Schauer, J. J.: Seasonal trends in the composition and sources of PM2.5 and carbonaceous aerosol in Tehran, Iran, Environ Pollut, 239, 69-81, 10.1016/j.envpol.2018.03.111, 2018.

Borgie, M., Ledoux, F., Dagher, Z., Verdin, A., Cazier, F., Courcot, L., Shirali, P., Greige-Gerges, H., and Courcot, D.: Chemical characteristics of PM 2.5–0.3 and PM 0.3 and consequence of a dust storm episode at an urban site in Lebanon, Atmospheric Research, 180, 274-286, 10.1016/j.atmosres.2016.06.001, 2016.

Cabaneros, S. M., Calautit, J. K., and Hughes, B. R.: A review of artificial neural network models for ambient air pollution prediction, Environmental Modelling & Software, 119, 285–304, 10.1016/j.envsoft.2019.06.014, 2019.

Cai, R., Yang, D., Ahonen, L. R., Shi, L., Korhonen, F., Ma, Y., Hao, J., Petäjä, T., Zheng, J., Kangasluoma, J., and Jiang, J.: Data inversion methods to determine sub-3 nm aerosol size distributions using the particle size magnifier, Atmos. Meas. Tech., 11, 4477-4491, 10.5194/amt-11-4477-2018, 2018.

Chaloulakou, A., Grivas, G., and Spyrellis, N.: Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment, J Air Waste Manag Assoc, 53, 1183-1190, 10.1080/10473289.2003.10466276, 2003.

Clifford, S., Low Choy, S., Hussein, T., Mengersen, K., and Morawska, L.: Using the Generalised Additive Model to model the particle number count of ultrafine particles, Atmospheric Environment, 45, 5934-5945, 10.1016/j.atmosenv.2011.05.004, 2011.

DeCarlo, P. F., Slowik, J. G., Worsnop, D. R., Davidovits, P., and Jimenez, J. L.: Particle morphology and density characterization by combined mobility and aerodynamic diameter measurements. Part 1: Theory, Aerosol Sci Tech, 38, 1185-1205, 10.1080/027868290903907, 2004.

Enting, I., and Newsam, G.: Atmospheric constituent inversion problems: Implications for baseline monitoring, Journal of Atmospheric Chemistry, 11, 69-87, 1990.

Fonseca, A. S., Viana, M., Perez, N., Alastuey, A., Querol, X., Kaminski, H., Todea, A. M., Monz, C., and Asbach, C.: Intercomparison of a portable and two stationary mobility particle sizers for nanoscale aerosol measurements, Aerosol Sci Tech, 50, 653-668, 10.1080/02786826.2016.1174329, 2016.

Freeman, B. S., Taylor, G., Gharabaghi, B., and Thé, J.: Forecasting air quality time series using deep learning, Journal of the Air & Waste Management Association, 68, 866–886, 10.1080/10962247.2018.1459956, 2018.

Fung, P. L., Zaidan, M. A., Timonen, H., Niemi, J. V., Kousa, A., Kuula, J., Luoma, K., Tarkoma, S., Petäjä, T., Kulmala, M., and Hussein, T.: Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration, Journal of Aerosol Science, 10.1016/j.jaerosci.2020.105694, 2020.

Gerling, L., Loschau, G., Wiedensohler, A., and Weber, S.: Statistical modelling of roadside and urban background ultrafine and accumulation mode particle number concentrations using generalized additive models, Sci Total Environ, 703, 134570, 10.1016/j.scitotenv.2019.134570, 2020.

Gherboudj, I., Beegum, S. N., and Ghedira, H.: Identifying natural dust source regions over the Middle-East and North-Africa: Estimation of dust emission potential, Earth-Sci Rev, 165, 342-355, 10.1016/j.earscirev.2016.12.010, 2017.

Goudarzi, G., Shirmardi, M., Naimabadi, A., Ghadiri, A., and Sajedifar, J.: Chemical and organic characteristics of PM2.5 particles and their in-vitro cytotoxic effects on lung cells: The Middle East dust storms in Ahvaz, Iran, Sci Total Environ, 655, 434-445, 10.1016/j.scitotenv.2018.11.153, 2019.

Gupta, R., and Xie, H.: Nanoparticles in Daily Life: Applications, Toxicity and Regulations, J Environ Pathol Toxicol Oncol, 37, 209-230, 10.1615/JEnvironPatholToxicolOncol.2018026009, 2018.

Hakala, S., Alghamdi, M. A., Paasonen, P., Vakkari, V., Khoder, M. I., Neitola, K., Dada, L., Abdelmaksoud, A. S., Al-Jeelani, H., Shabbaj, I. I., Almehmadi, F. M., Sundström, A. M., Lihavainen, H., Kerminen, V. M., Kontkanen, J., Kulmala, M., Hussein, T., and Hyvärinen, A. P.: New particle formation, growth and apparent shrinkage at a rural background site in western Saudi Arabia, Atmos. Chem. Phys., 19, 10537-10555, 10.5194/acp-19-10537-2019, 2019.

Hellén, H., Kangas, L., Kousa, A., Vestenius, M., Teinilä, K., Karppinen, A., Kukkonen, J., and Niemi, J. V.: Evaluation of the impact of wood combustion on benzo[a]pyrene (BaP) concentrations; ambient measurements and dispersion modeling in Helsinki, Finland, Atmospheric Chemistry and Physics, 17, 3475-3487, 10.5194/acp-17-3475-2017, 2017.

Hussein, T., Dal Maso, M., Petäjä, T., Koponen, I. K., Paatero, P., Aalto, P. P., Hämeri, K., and Kulmala, M.: Evaluation of an automatic algorithm for fitting the particle number size distributions, Boreal Environment Research, 10, 337–355, 2005.

Hussein, T., Kukkonen, J., Korhonen, H., Pohjola, M., Pirjola, L., Wraith, D., Harkonen, J., Teinila, K., Koponen, I. K., Karppinen, A., Hillamo, R., and Kulmala, M.: Evaluation and modeling of the size fractionated aerosol particle number concentration measurements nearby a major road in Helsinki – Part II: Aerosol measurements within the SAPPHIRE project, Atmospheric Chemistry and Physics, 7, 4081-4094, 10.5194/acp-7-4081-2007, 2007.

Hussein, T., Dada, L., Hakala, S., Petäjä, T., and Kulmala, M.: Urban Aerosol Particle Size Characterization in Eastern Mediterranean Conditions, Atmosphere, 10, 10.3390/atmos10110710, 2019.

Hussein, T., Atashi, N., Sogacheva, L., Hakala, S., Dada, L., Petäjä, T., and Kulmala, M.: Characterization of Urban New Particle Formation in Amman—Jordan, Atmosphere, 11, 10.3390/atmos11010079, 2020.

Kandlikar, M., and Ramachandran, G.: Inverse methods for analysing aerosol spectrometer measurements: a critical review, Journal of Aerosol Science, 30, 413-437, 1999.

Kannosto, J., Virtanen, A., Lemmetty, M., Mäkelä, J. M., Keskinen, J., Junninen, H., Hussein, T., Aalto, P., and Kulmala, M.: Mode resolved density of atmospheric aerosol particles, Atmos. Chem. Phys., 8, 5327-5337, 10.5194/acp-8-5327-2008, 2008.

Kerminen, V. M., Paramonov, M., Anttila, T., Riipinen, I., Fountoukis, C., Korhonen, H., Asmi, E., Laakso, L., Lihavainen, H., Swietlicki, E., Svenningsson, B., Asmi, A., Pandis, S. N., Kulmala, M., and Petaja, T.: Cloud condensation nuclei production associated with atmospheric nucleation: a synthesis based on existing literature and new results, Atmospheric Chemistry and Physics, 12, 12037-12059, 10.5194/acp-12-12037-2012, 2012.

Kerminen, V. M., Chen, X. M., Vakkari, V., Petaja, T., Kulmala, M., and Bianchi, F.: Atmospheric new particle formation and growth: review of field observations, Environ Res Lett, 13, 10.1088/1748-9326/aadf3c, 2018.

Kok, J. F., Ridley, D. A., Zhou, Q., Miller, R. L., Zhao, C., Heald, C. L., Ward, D. S., Albani, S., and Haustein, K.: Smaller desert dust cooling effect estimated from analysis of dust size and abundance, Nat Geosci, 10, 274-278, 10.1038/Ngeo2912, 2017.

Kreyling, W. G., Semmler, M., and Moller, W.: Dosimetry and toxicology of ultrafine particles, J Aerosol Med, 17, 140-152, 10.1089/0894268041457147, 2004.

Kulkarni, P., Baron, P. A., and Willeke, K.: Aerosol measurement: principles, techniques, and applications, John Wiley & Sons, 2011.

Kulmala, M., Vehkamaki, H., Petaja, T., Dal Maso, M., Lauri, A., Kerminen, V. M., Birmili, W., and McMurry, P. H.: Formation and growth rates of ultrafine atmospheric particles: a review of observations, Journal of Aerosol Science, 35, 143-176, 10.1016/j.jaerosci.2003.10.003, 2004.

Laakso, L., Hussein, T., Aarnio, P., Komppula, M., Hiltunen, V., Viisanen, Y., and Kulmala, M.: Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland, Atmospheric Environment, 37, 2629-2641, 10.1016/S1352-2310(03)00206-1, 2003.

Lehtipalo, K., Leppa, J., Kontkanen, J., Kangasluoma, J., Franchin, A., Wimnner, D., Schobesberger, S., Junninen, H., Petaja, T., and Sipila, M. J. B. E. R.: Methods for determining particle size distribution and growth rates between 1 and 3 nm using the Particle Size Magnifier, 2014.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367-371, 10.1038/nature15371, 2015.

Lewis, A. C., Lee, J. D., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., Smith, K. R., Buckley, J. W., Ellis, M., Gillot, S. R., and White, A.: Evaluating the performance of low cost chemical sensors for air pollution research, Faraday Discuss, 189, 85-103, 10.1039/c5fd00201j, 2016.

Liu, Z. R., Hu, B., Liu, Q., Sun, Y., and Wang, Y. S.: Source apportionment of urban fine particle number concentration during summertime in Beijing, Atmospheric Environment, 96, 359-369, 10.1016/j.atmosenv.2014.06.055, 2014.

Londahl, J., Moller, W., Pagels, J. H., Kreyling, W. G., Swietlicki, E., and Schmid, O.: Measurement techniques for respiratory tract deposition of airborne nanoparticles: a critical review, J Aerosol Med Pulm Drug Deliv, 27, 229-254, 10.1089/jamp.2013.1044, 2014.

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y. T., and Rahmati, M.: Air pollution prediction by using an artificial neural network model, Clean Technologies and Environmental Policy, 21, 1341–1352, 10.1007/s10098-019-01709-w, 2019.

Markowski, G. R.: Improving Twomey's algorithm for inversion of aerosol measurement data, Aerosol Sci Tech, 7, 127-141, 1987.

Mølgaard, B., Birmili, W., Clifford, S., Massling, A., Eleftheriadis, K., Norman, M., Vratolis, S., Wehner, B., Corander, J., Hämeri, K., and Hussein, T.: Evaluation of a statistical forecast model for size-fractionated urban particle number concentrations using data from five European cities, Journal of Aerosol Science, 66, 96-110, 10.1016/j.jaerosci.2013.08.012, 2013.

Morawska, L., Ristovski, Z., Jayaratne, E. R., Keogh, D. U., and Ling, X.: Ambient nano and ultrafine particles from motor vehicle emissions: Characteristics, ambient processing and implications on human exposure, Atmospheric Environment, 42, 8113-8138, 10.1016/j.atmosenv.2008.07.050, 2008.

Ohlwein, S., Kappeler, R., Joss, M. K., Kunzli, N., and Hoffmann, B.: Health effects of ultrafine particles: a systematic literature review update of epidemiological evidence, Int J Public Health, 64, 547-559, 10.1007/s00038-019-01202-7, 2019.

Popoola, O. A. M., Stewart, G. B., Mead, M. I., and Jones, R. L.: Development of a baseline-temperature correction methodology for electrochemical sensors and its implications for long-term stability, Atmospheric Environment, 147, 330-343, 10.1016/j.atmosenv.2016.10.024, 2016.

Rahman, M. M., Mazaheri, M., Clifford, S., and Morawska, L.: Estimate of main local sources to ambient ultrafine particle number concentrations in an urban area, Atmospheric Research, 194, 178-189, 10.1016/j.atmosres.2017.04.036, 2017.

Reggente, M., Peters, J., Theunis, J., Van Poppel, M., Rademaker, M., Kumar, P., and De Baets, B.: Prediction of ultrafine particle number concentrations in urban environments by means of Gaussian process regression based on measurements of oxides of nitrogen, Environmental Modelling & Software, 61, 135-150, 10.1016/j.envsoft.2014.07.012, 2014.

Rönkkö, T., Kuuluvainen, H., Karjalainen, P., Keskinen, J., Hillamo, R., Niemi, J. V., Pirjola, L., Timonen, H. J., Saarikoski, S., Saukko, E., Jarvinen, A., Silvennoinen, H., Rostedt, A., Olin, M., Yli-Ojanpera, J., Nousiainene, P., Kousa, A., and Dal Maso, M.: Traffic is a major source of atmospheric nanocluster aerosol, P Natl Acad Sci USA, 114, 7549-7554, 10.1073/pnas.1700830114, 2017.

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence, 1, 206–215, 10.1038/s42256-019-0048-x, 2019.

Sabaliauskas, K., Jeong, C. H., Yao, X. H., Jun, Y. S., Jadidian, P., and Evans, G. J.: Five-year roadside measurements of ultrafine particles in a major Canadian city, Atmospheric Environment, 49, 245-256, 10.1016/j.atmosenv.2011.11.052, 2012.

Spinazzè, A., Fanti, G., Borghi, F., Del Buono, L., Campagnolo, D., Rovelli, S., Cattaneo, A., and Cavallo, D. M.: Field comparison of instruments for exposure assessment of airborne ultrafine particles and particulate matter, Atmospheric Environment, 154, 274-284, 10.1016/j.atmosenv.2017.01.054, 2017.

Stolzenburg, M. R., and McMurry, P. H.: Method to assess performance of scanning mobility particle sizer (SMPS) instruments and software, Aerosol Sci Tech, 52, 609-613, 10.1080/02786826.2018.1455962, 2018.

Teinilä, K., Aurela, M., Niemi, J. V., Kousa, A., Petäjä, T., Järvi, L., Hillamo, R., Kangas, L., Saarikoski, S., and Timonen, H.: Concentration variation of gaseous and particulate pollutants in the Helsinki city centre — observations from a two-year campaign from 2013–2015, Boreal Environment Research, 24, 115–136, 2019.

Tritscher, T., Beeston, M., Zerrath, A. F., Elzey, S., Krinke, T. J., Filimundi, E., and Bischof, O. F.: NanoScan SMPS - A Novel, Portable Nanoparticle Sizing and Counting Instrument, J Phys Conf Ser, 429, 10.1088/1742-6596/429/1/012061, 2013.

Tritscher, T., Koched, A., Han, H. S., Filimundi, E., Johnson, T., Elzey, S., Avenido, A., Kykal, C., and Bischof, O. F.: Multi-Instrument Manager Tool for Data Acquisition and Merging of Optical and Electrical Mobility Size Distributions, 4th International Conference on Safe Production and Use of Nanomaterials (Nanosafe2014), 617, 10.1088/1742-6596/617/1/012013, 2015.

Viskari, T., Asmi, E., Kolmonen, P., Vuollekoski, H., Petaja, T., and Jarvinen, H.: Estimation of aerosol particle number distributions with Kalman Filtering - Part 1: Theory, general aspects and statistical validity, Atmospheric Chemistry and Physics, 12, 11767-11779, 10.5194/acp-12-11767-2012, 2012.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., Fjäraa, A. M., Asmi, E., Sellegri, K., Depuy, R., Venzac, H., Villani, P., Laj, P., Aalto, P., Ogren, J. A., Swietlicki, E., Williams, P., Roldin, P., Quincey, P., Hüglin, C., Fierz-Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Grüning, C., Faloon, K., Beddows, D., Harrison, R., Monahan, C., Jennings, S. G., O'Dowd, C. D., Marinoni, A., Horn, H. G., Keck, L., Jiang, J., Scheckman, J., McMurry, P. H., Deng, Z., Zhao, C. S., Moerman, M., Henzing, B., de Leeuw, G., Löschau, G., and Bastian, S.: Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions, Atmos. Meas. Tech., 5, 657-685, 10.5194/amt-5-657-2012, 2012.

Population growth (annual %): https://data.worldbank.org/indicator/SP.POP.GROW, access: 06-10, 2019.

World Health Organisation: World health statistics 2019: Monitoring health for the SDGs, sustainable development goals, World Health Organisation, https://apps.who.int/iris/handle/10665/324835, 2019.

Xing, Y. F., Xu, Y. H., Shi, M. H., and Lian, Y. X.: The impact of PM2.5 on the human respiratory system, J Thorac Dis, 8, E69-74, 10.3978/j.issn.2072-1439.2016.01.19, 2016.

Zaidan, M. A., Canova, F. F., Laurson, L., and Foster, A. S.: Mixture of Clustered Bayesian Neural Networks for Modeling Friction Processes at the Nanoscale, J Chem Theory Comput, 13, 3-8, 10.1021/acs.jctc.6b00830, 2017.

Zaidan, M. A., Wraith, D., Boor, B. E., and Hussein, T.: Bayesian Proxy Modelling for Estimating Black Carbon Concentrations using White-Box and Black-Box Models, Applied Sciences, 9, 10.3390/app9224976, 2019.

Zaidan, M. A., Surakhi, O., Fung, P. L., and Hussein, T.: Sensitivity Analysis for Predicting Sub-Micron Aerosol Concentrations Based on Meteorological Parameters, Sensors (Basel), 20, 10.3390/s20102876, 2020.

Zhou, Y., Dada, L., Liu, Y., Fu, Y., Kangasluoma, J., Chan, T., Yan, C., Chu, B., Daellenbach, K. R., Bianchi, F. J. A. C., and Physics: Variation of size-segregated particle number concentrations in wintertime Beijing, 20, 1201-1216, 2020.

https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021

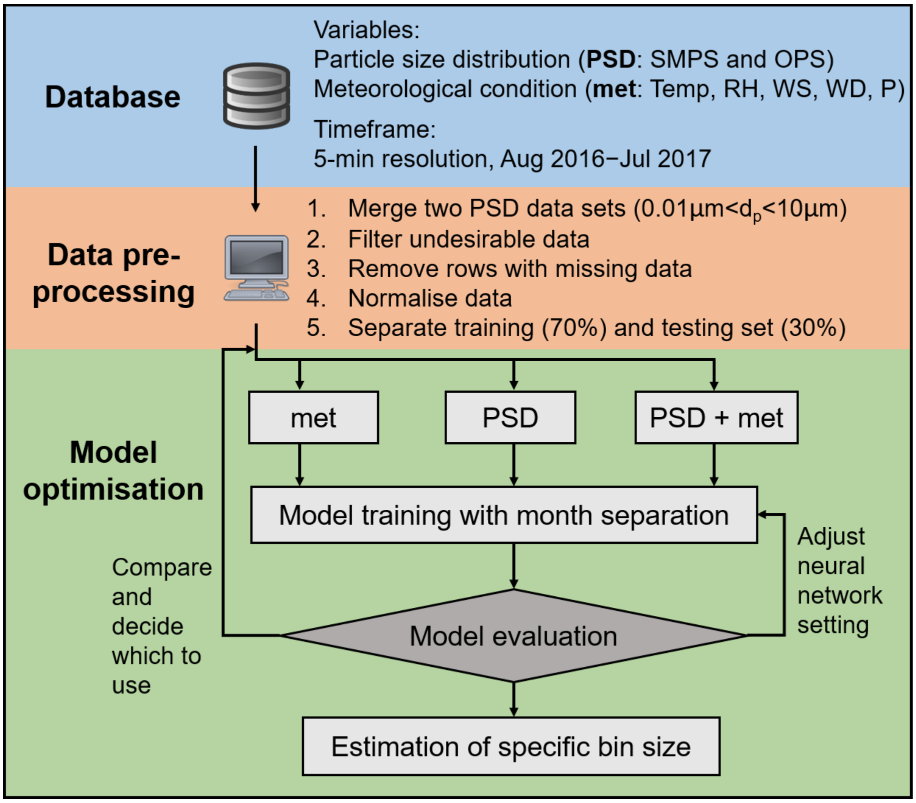

Database
Variables:
Particle size distribution (**PSD**: SMPS and OPS)
Meteorological condition (**met**: Temp, RH, WS, WD, P)
Timeframe:
5-min resolution, Aug 2016−Jul 2017
Data pre-processing
1. Merge two PSD data sets (0.01µm<d_p<10µm)
2. Filter undesirable data
3. Remove rows with missing data
4. Normalise data
5. Separate training (70%) and testing set (30%)
Model optimisation
met
PSD
PSD + met
Model training with month separation
Adjust neural network setting
Compare and decide which to use
Model evaluation
Estimation of specific bin size

Figure 1. The block diagram describing the methodology of the model.

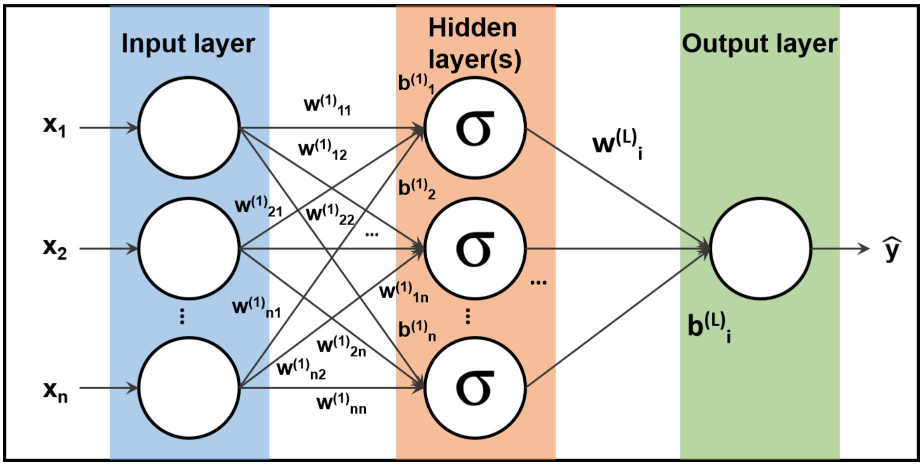


Figure 2. Schematic diagram of a neural network with one hidden layer of sigmoid activation function.

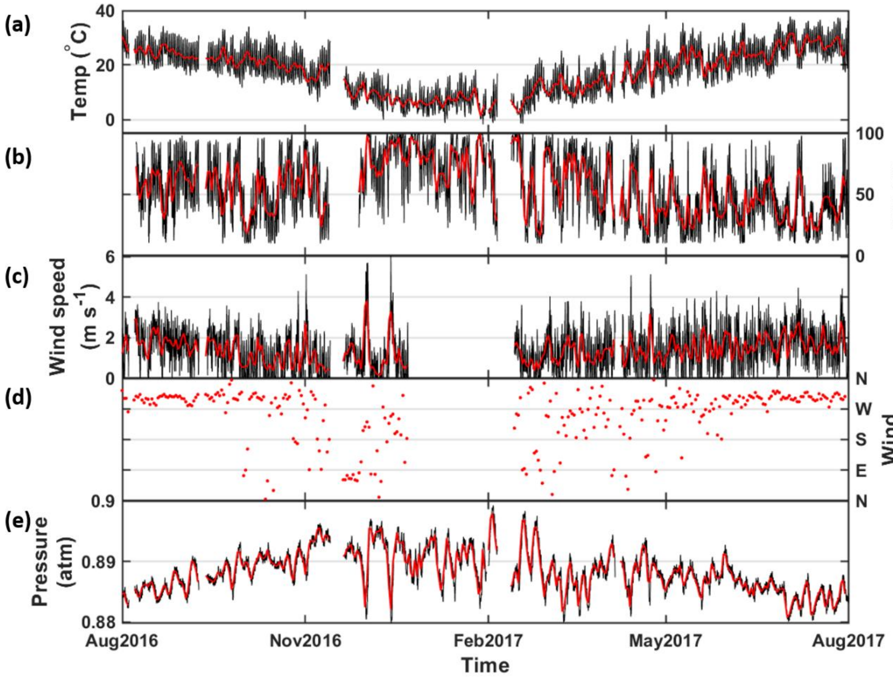


Figure 3. Timeseries of meteorological conditions during the measurement period Aug 2016–Jul 2017. (a–e) denotes temperature, relative humidity, wind speed, wind direction and air pressure, respectively. Black and red represent hourly and daily averaged data, respectively.

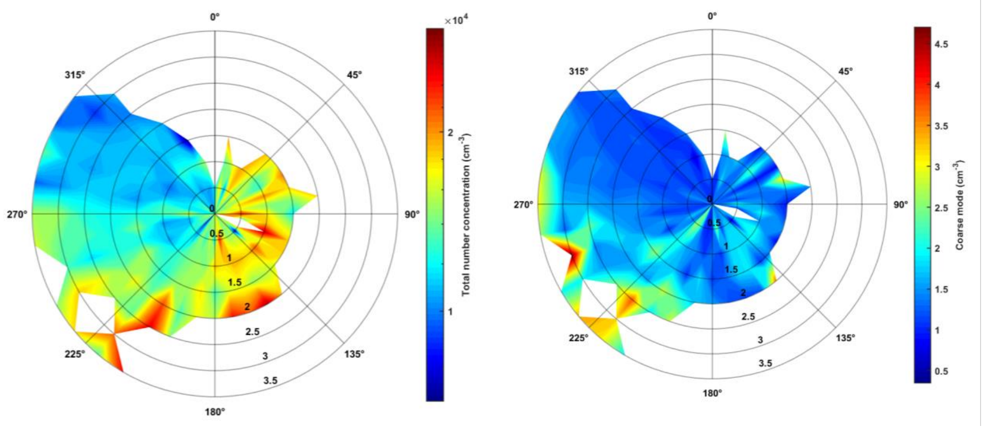


Figure 4. Windrose diagram of total particle number concentration at different direction (in theta axis) and different wind speed (in radical axis). Wind direction and wind speed data are grouped in every 10^o and 0.5 m s^{-1}. Warmer color represent higher total particle number concentration. (a) total number concentration, log scale; (b) coarse mode, linear scale. Note the color scales are different.

https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021

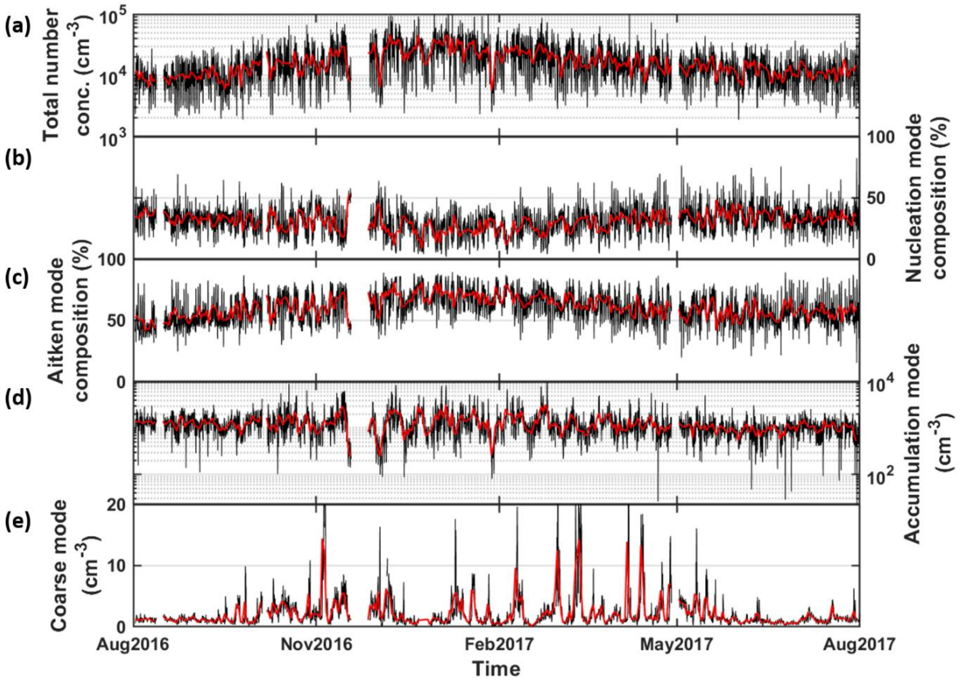


Figure 5. Timeseries of total particle number concentration (in cm^{-3}) of 0.01–10µm in (a). (b–c) indicate the composition in percentage of nucleation mode and Aitken mode, respectively. (d–e) show the number concentration in accumulation mode and coarse mode, respectively. Black and red represent hourly and daily averaged data, respectively.

https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021

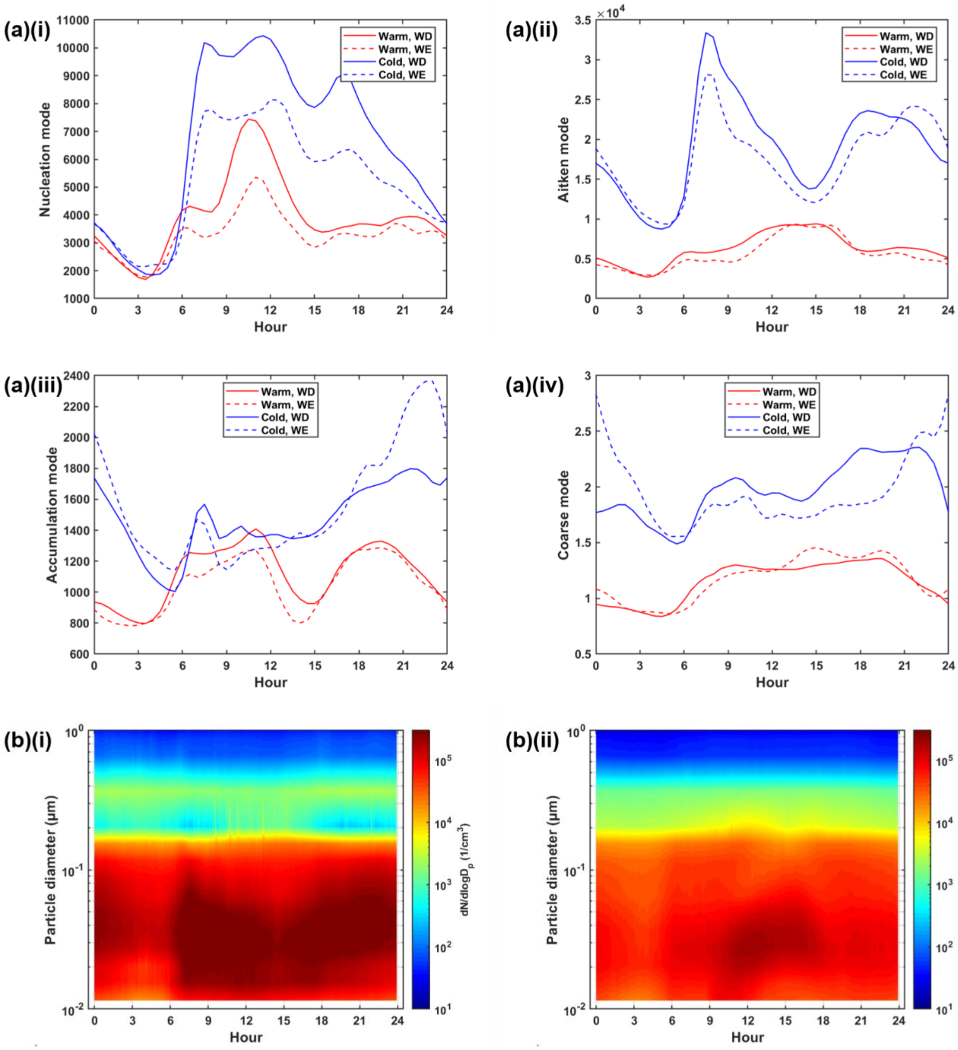


Figure 6. (a) Diurnal cycle of the (i) nucleation mode, (ii) Aitken mode, (iii) accumulation mode and (iv) coarse mode in warm (red) and cold months (blue) during workdays (solid) and weekends (dashed). (b) Particle size distribution in (i) cold and (ii) warm months, coloured by particle number concentration (cm^{-3}).

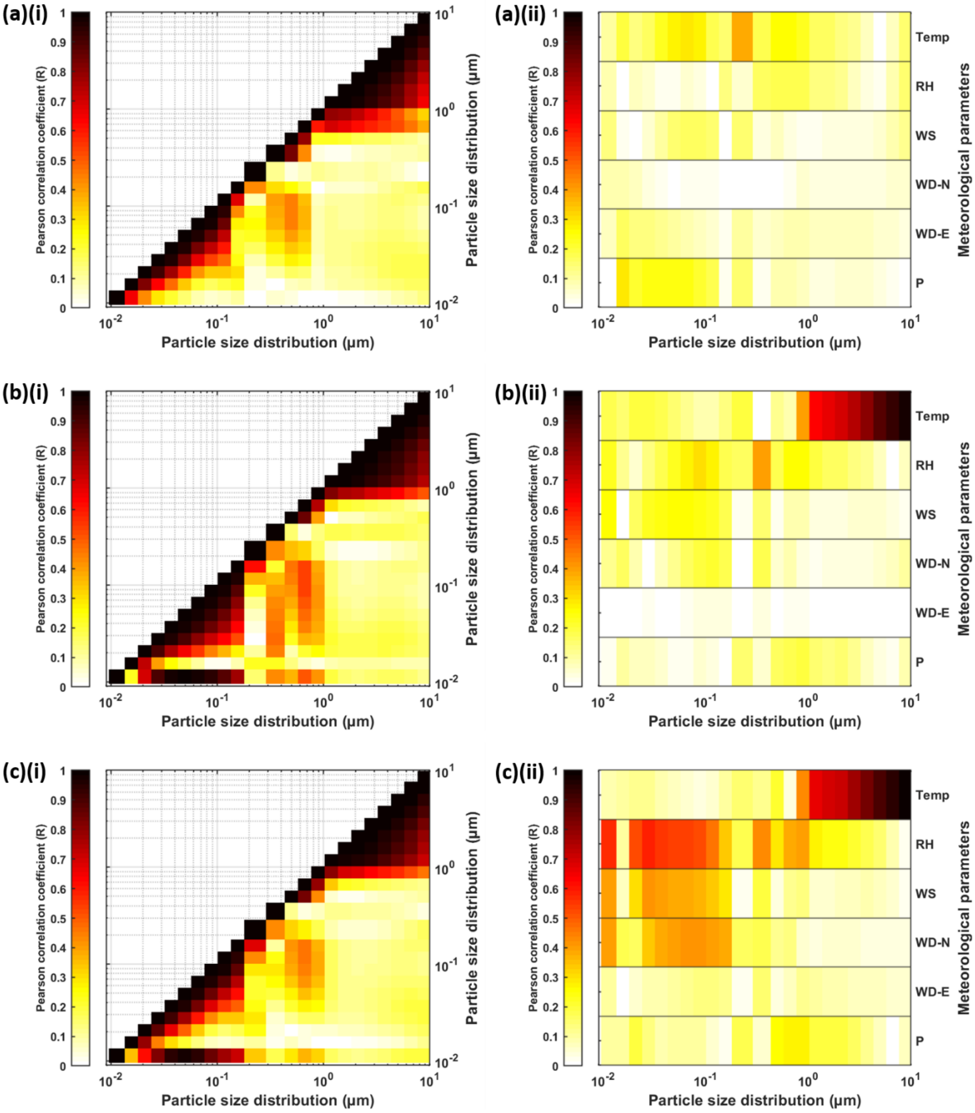
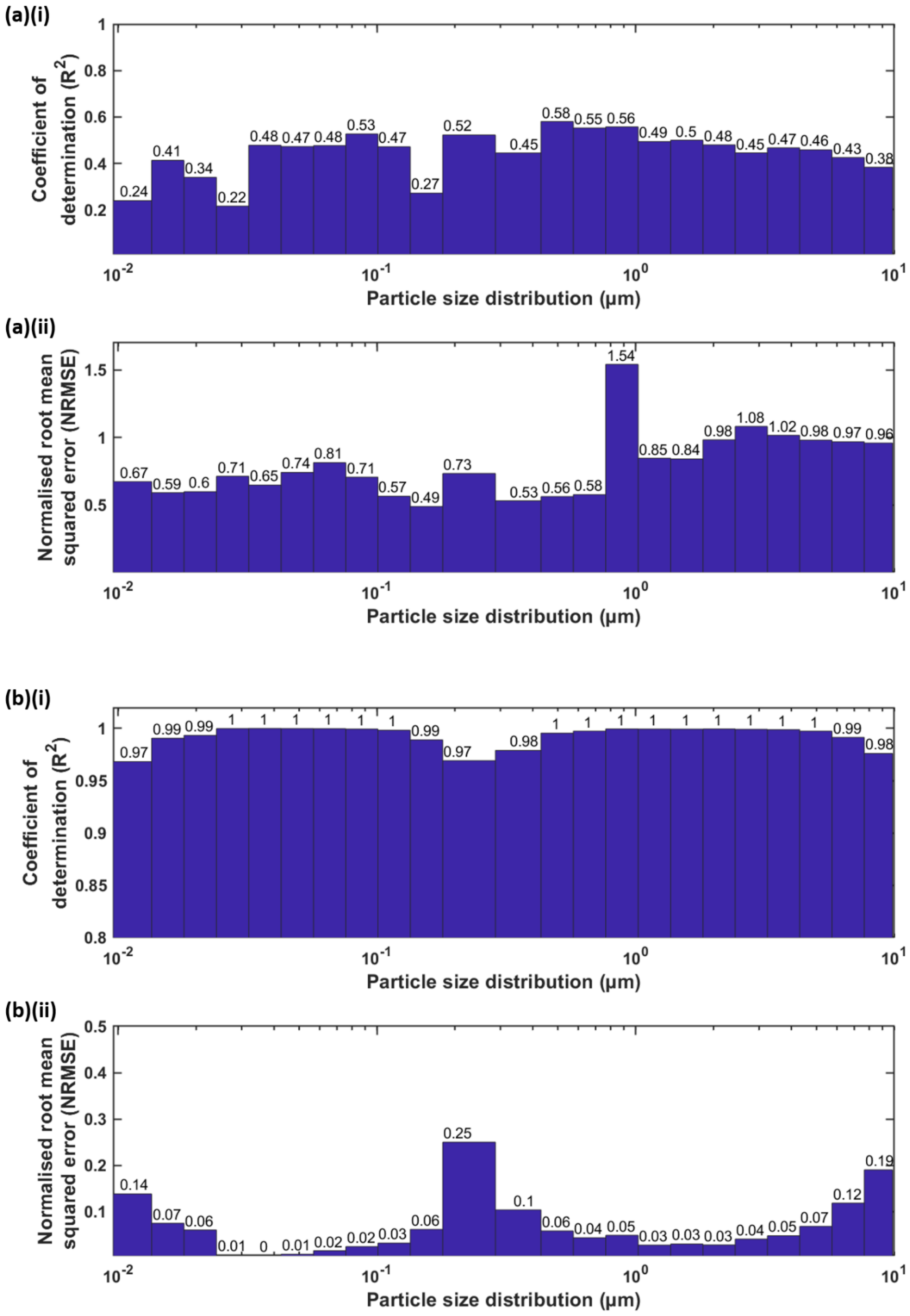


Figure 7. Matrix plots showing the Pearson correlation coefficient (R) of particle size distribution of (a) 5-min, (b) hourly, (c) daily averaging with (i) particle size distribution itself and (ii) meteorological parameters. Darker colour represents a higher correlation.

https://doi.org/10.5194/amt-2021-37
Preprint. Discussion started: 18 March 2021



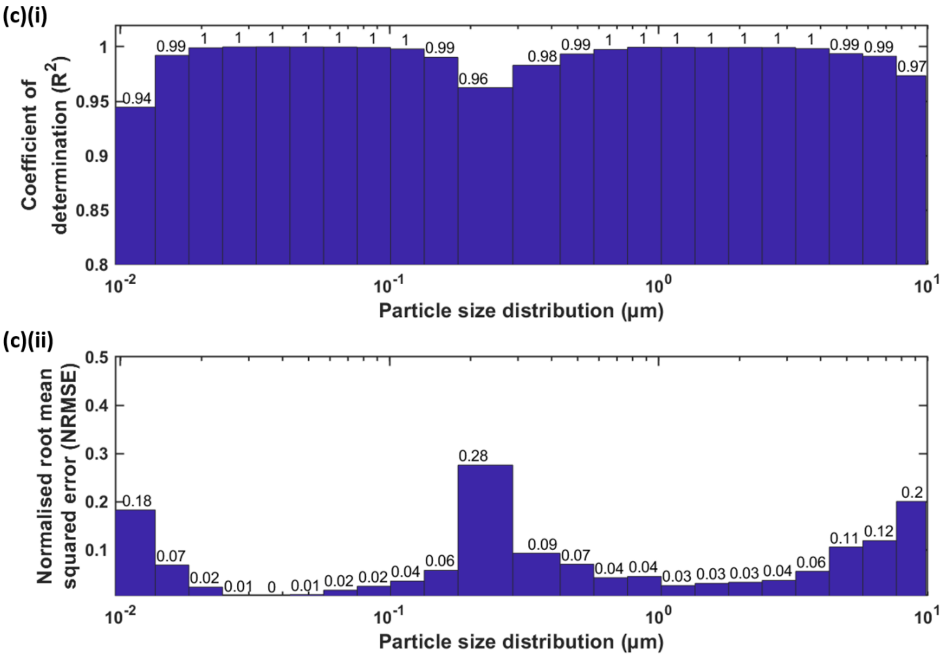


Figure 8. Bar chart showing the model evaluation of model with (a) only meteorological parameters, (b) particle size distribution itself, (c) both particle size distribution and meteorological parameters as inputs. The model evaluation metrics include (i) coefficient of determination (R^2) and (ii) normalised root mean squared error (NRMSE).

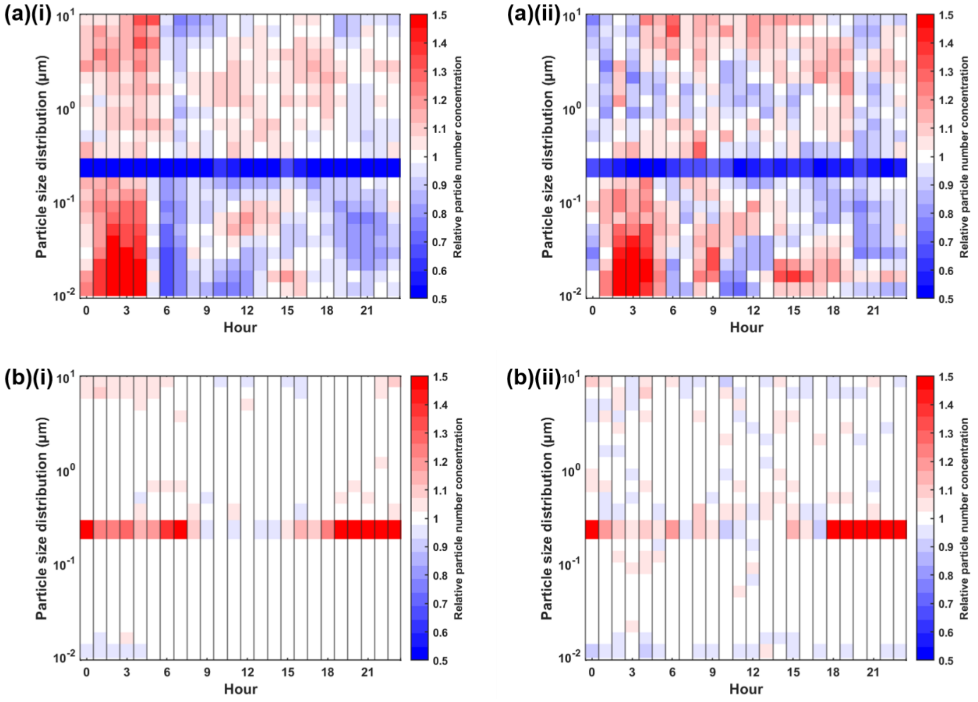


Figure 9. Heatmap showing the hourly median relative particle number concentration of the models with (a) meteorological parameters and (b) particle size distribution as input across different hours of a day (i) in workdays and (ii) in weekends. The relative particle number concentration is defined as modelled concentration with respect to measured concentration. Red colour show overestimation while blue show underestimation.

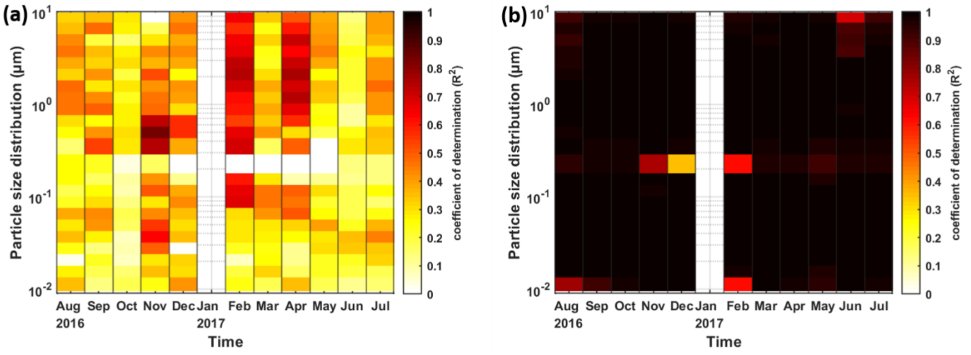


Figure 10. Heatmap showing the coefficient of determination (R^2) of the models with (a) meteorological parameters and (b) particle size distribution as input for different months at different size bins. Darker colour represents a higher R^2.

Table 1. Table showing the descriptive statistics (in cm^{-3}) of total number concentration, nucleation mode, Aitken mode, accumulation mode and coarse mode. The statistical values include mean, standard deviation, and percentile (10%, 25%, 50%, 75% and 90%).

	Mean	std	10%	25%	50%	75%	90%
Total (×10^{4})	1.70	1.26	0.57	0.85	1.35	2.16	3.31
Nucleation (×10^{4})	0.48	0.32	0.16	0.26	0.41	0.63	0.90
Aitken (×10^{4})	1.09	1.01	0.29	0.45	0.77	1.37	2.35
Accumulation (×10^{4})	0.13	0.08	0.05	0.08	0.11	0.15	0.21
Coarse	2.13	2.80	0.55	0.84	1.29	2.33	4.3

Table 2. Table showing the best configuration in the form of (the number of layers; the number of neurons) for the model by meteorological parameters and the number concentration at the other size bins. Mean absolute error (MAE, in cm^{-3}), coefficient of determination (R^{2}) and normalised root-mean-square error (NRMSE) are listed for different size bins on each row. The last row concludes the overall selection of the model with the best configuration and its corresponding evaluation metrics.

Particle	Approach 1 (met)				Approach 2 (PSD)			
size (μm)	Best setting	MAE (cm^{-3})	R^{2}	NRMSE	Best setting	MAE (cm^{-3})	R^{2}	NRMSE
0.012	2; 10	2638.6	0.1996	0.6918	2; 10	333.7442	0.99	0.1077
0.015	2; 15	4853.0	0.4237	0.5868	2; 8	215.8006	1.00	0.0310
0.021	2; 15	6119.1	0.3774	0.5831	2; 15	97.8408	1.00	0.0136
0.027	2; 15	8469.2	0.4072	0.6210	1; 25	34.0126	1.00	0.0032
0.037	2; 20	8235.7	0.4568	0.6619	2; 15	26.2854	1.00	0.0024
0.049	2; 15	6608.3	0.4778	0.7389	2; 25	33.7074	1.00	0.0049
0.066	2; 15	4688.5	0.4613	0.8266	2; 10	56.7074	1.00	0.0125
0.088	2; 15	3041.6	0.5207	0.7114	2; 4	66.1841	1.00	0.0183
0.12	2; 15	1806.3	0.5193	0.5398	2; 8	63.1301	1.00	0.0210
0.15	2; 10	917.25	0.2836	0.4865	2; 15	72.4539	0.99	0.0515
0.21	2; 6	326.66	0.5536	0.7101	2; 8	114.3451	0.91	0.3142
0.37	2; 10	95.84	0.4297	0.5396	2; 20	12.8995	0.99	0.0723
0.49	2; 15	12.06	0.5025	0.6138	2; 25	0.9630	1.00	0.0427
0.66	2; 15	3.03	0.5824	0.5580	2; 15	0.1995	1.00	0.0290
0.88	2; 15	5.65	0.6190	1.4301	2; 10	0.2202	1.00	0.0398
1.17	2; 15	1.43	0.5331	0.8134	2; 8	0.0680	1.00	0.0257
1.56	2; 20	1.44	0.5384	0.8088	2; 8	0.0816	1.00	0.0312
2.08	2; 15	1.84	0.4885	0.9748	2; 8	0.0825	1.00	0.0278
2.77	2; 15	1.02	0.4352	1.0925	1; 4	0.0573	1.00	0.0372
3.70	2; 15	0.52	0.4076	1.0719	1; 8	0.0329	1.00	0.0455
4.92	2; 15	0.28	0.4427	0.9955	1; 4	0.0254	1.00	0.0681
6.56	2; 9	0.11	0.4231	0.9710	1; 6	0.0206	0.99	0.1252
8.75	2; 10	0.060	0.3903	0.9546	2; 6	0.0169	0.98	0.1980
overall	2; 15	2118.577	0.67	1.1324	2; 10	76.6329	0.9989	0.0671