Atmos. Chem. Phys., 18, 12699–12714, 2018
https://doi.org/10.5194/acp-18-12699-2018

Atmospheric Chemistry and Physics
Open Access
EGU

Exploring non-linear associations between atmospheric new-particle formation and ambient variables: a mutual information approach

Martha A. Zaidan[1,2,3], **Ville Haapasilta**[3], **Rishi Relan**[4], **Pauli Paasonen**[1], **Veli-Matti Kerminen**[1], **Heikki Junninen**[1,5], **Markku Kulmala**[1,6], **and Adam S. Foster**[3,7,8]

[1]Institute for Atmospheric and Earth System Research/Physics, Helsinki University, 00560 Helsinki, Finland
[2]Aalto Science Institute, School of Science, Aalto University, 00076 Espoo, Finland
[3]Department of Applied Physics, Aalto University, 00076 Espoo, Finland
[4]Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[5]Institute of Physics, University of Tartu, Ülikooli 18, 50090 Tartu, Estonia
[6]Aerosol and Haze Laboratory, Beijing University of Chemical Technology, 100096 Beijing, China
[7]WPI Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan
[8]Graduate School Materials Science in Mainz, Staudinger Weg 9, 55128 Mainz, Germany

Correspondence: Martha A. Zaidan (martha.zaidan@helsinki.fi)

Received: 12 February 2018 – Discussion started: 20 April 2018
Revised: 6 August 2018 – Accepted: 15 August 2018 – Published: 3 September 2018

Abstract. Atmospheric new-particle formation (NPF) is a very non-linear process that includes atmospheric chemistry of precursors and clustering physics as well as subsequent growth before NPF can be observed. Thanks to ongoing efforts, now there exists a tremendous amount of atmospheric data, obtained through continuous measurements directly from the atmosphere. This fact makes the analysis by human brains difficult but, on the other hand, enables the usage of modern data science techniques. Here, we calculate and explore the mutual information (MI) between observed NPF events (measured at Hyytiälä, Finland) and a wide variety of simultaneously monitored ambient variables: trace gas and aerosol particle concentrations, meteorology, radiation and a few derived quantities. The purpose of the investigations is to identify key factors contributing to the NPF. The applied mutual information method finds that the formation events are strongly linked to sulfuric acid concentration and water content, ultraviolet radiation, condensation sink (CS) and temperature. Previously, these quantities have been well-established to be important players in the phenomenon via dedicated field, laboratory and theoretical research. The novelty of this work is to demonstrate that the same results are now obtained by a data analysis method which operates without supervision and without the need of understanding the physics deeply. This suggests that the method is suitable to be implemented widely in the atmospheric field to discover other interesting phenomena and their relevant variables.

1 Introduction

New-particle formation (NPF) is an important source of aerosol particles and cloud condensation nuclei (CCN) and in a vast number of atmospheric environments ranging from remote continental areas to heavily polluted urban centres (Kulmala et al., 2004; Dunne et al., 2016; Wang et al., 2017). The occurrence and strength of NPF and its influence on the CCN budget in different atmospheric environments depends on a delicate balance between the factors that favour NPF and subsequent particle growth and the factors that suppress these processes (Kerminen and Kulmala, 2002; Pierce and Adams, 2007; Westervelt et al., 2014; Kulmala et al., 2017). As a result, researchers have not managed to find a general framework, or formulae, on how to relate atmospheric NPF to the concentrations of various trace gases, meteorological quantities and radiation parameters.

Based on data from field measurements, several studies investigated the relations between NPF and meteorological conditions (Nilsson et al., 2001) and various chemical compounds (Bonn and Moortgat, 2003; Kulmala et al., 2004; Almeida et al., 2013; Nieminen et al., 2014). Such studies have found the ideal conditions for NPF events to consist of low atmospheric water content, low preexisting particle concentration and high solar radiation (Boy and Kulmala, 2002). In addition, sulfuric acid is believed to be the single most important compound to participate in the atmospheric NPF (Kerminen et al., 2010; Sipilä et al., 2010; Petäjä et al., 2011; Nieminen et al., 2014).

Due to the practical limitations, the measurement campaigns typically last from weeks to months and they often have a dedicated focus. On the one hand, such an approach enables a very detailed inspection for a somewhat narrower scope, but, on the other hand, there is a risk of overlooking important processes falling outside the chosen, predetermined scope. One way to circumvent this issue is to have long-term continuous measurements of a wide variety of atmospheric variables. Nowadays there is more and more focus on continuous observations as described by Kulmala (2018). However, such enterprises then open a new set problems: how to analyse all the collected data? It is clear that techniques offered by the modern data science, such as data mining and machine learning, should be consulted.

Previously, Mikkonen et al. (2006) studied the effects of gas and meteorological parameters as well as aerosol size distribution to nucleation events. The used data were measured in the Po Valley, Italy, for about 3 years (2002–2005). In this case, they used a discriminant analysis method where relative humidity (RH), ozone and radiation are found to give the best classification performance. Next, similar atmospheric variables were also included in their further study (Mikkonen et al., 2011). The used data were measured from three polluted sites, which are the Po Valley, Italy; and Melpitz and Hohenpeissenberg, Germany. In this study, they applied a multivariate non-linear mixed effects model to examine the variables affecting the number concentration of Aitken particles (50 nm). They also found that relative humidity and ozone give the best predictor variables. In addition, the model indicated that the temperature, condensation sink (CS), and concentrations of sulfuric dioxide (SO_2) and nitrogen dioxide (NO_2) influence NPF as well as the number concentration of Aitken mode particles.

In order to understand the effects of atmospheric variables to NPF in Hyytiälä, Finland, a comprehensive study was done by Hyvönen et al. (2005). They utilized two main types of data mining methods on 8 years of continuous measurements of 80 variables. Their first method was based on unsupervised K-means clustering. The first method demonstrated that the relative humidity, global radiation and sensible heat have data separation power and correlate with NPF. In addition to those, their results indicated that ozone (O_3) and carbon dioxide (CO_2) concentrations might also correlate with NPF. The second method was based on a supervised learning classification. Several machine learning models (such as linear discriminant analysis, support vector machine and logistic regression) were set up to perform a classification task for each day as an event or a non-event day. The goal was not to separate event days from non-event days, but to understand which atmospheric variables should be used to clearly separate the two groups. In this case, the mean and standard deviation of atmospheric variables were calculated as the input, whereas the aerosol particle formation event and non-event days database was used as the output. Due to the initial model's random parameters, the models were run 1000 times using different training and test sets to ensure the result stability. The selected models used a pair and triplet combination of atmospheric variables. The models were ranked based on the classification performance and the best model was used to evaluate all pair and triplet combinations of the atmospheric variables. In this case, the supervised classification models found that the best pair of atmospheric variables to classify events–non-events is condensation sink and relative humidity. The latter was also found through the clustering method. The results of Hyvönen et al. (2005) support some earlier conclusions from Boy and Kulmala (2002) stating that NPF events are largely explained by three parameters: temperature, the atmospheric water content and radiation. However, they did not find significant correlations between NPF and radiation variables as suggested by the aforementioned studies.

The previously used data mining approaches are mostly based on classification methods. Although these methods seem to be suitable tools for finding correlation between variables in complex systems, the used implementation may not be always effective for this case. The first reason concerns the used features, such as mean and standard deviations. This practice compresses the measurement data into a single quantity for each day, which may potentially lead to information loss in the data. Secondly, the implementation procedure is computationally expensive. This requires the exploration of all possible models and variable combinations to find the best pairs. The models also need to be run multiple times to ensure their stability.

To overcome the above-mentioned issues, we propose here an alternative method – based on information theory – to be used in atmospheric data analysis. Mutual information (MI), one of the many information quantities, measures the amount of information that can be obtained about one random variable by observing another one. In this paper, MI is first introduced and then used to find the maximal amount of shared information between atmospheric variables and NPF. In other words, the goal is to find the most relevant atmospheric variables in relation to NPF events using a data-driven information theoretic method based on the data set measured at Hyytiälä, Finland.

2 Atmospheric database measured at the SMEAR II station in Hyytiälä, Finland

In this study, we utilize the data measured during the years 1996–2014 at the Station for Measuring Forest Ecosystem-Atmosphere Relations (SMEAR) II station in Hyytiälä, Finland, operated by Helsinki University (SMEAR website, 2017).

2.1 Sampling site

The SMEAR II station is located in Hyytiälä forestry field station in southern Finland (61°51′ N, 24°17′ E; 181 m above sea level), about 220 km northwest of Helsinki. It also lies between two large cities, Tampere and Jyväskylä, that are about 60 and 90 km from the measurement site, respectively. Homogeneous 55-year-old (in 2017) scots-pine-dominated forests surround the station. SMEAR II is classified as a rural background site considering the levels of air pollutants, shown by for example submicron aerosol number size distributions (Asmi et al., 2011a; Nieminen et al., 2014).

The SMEAR II station has been established for multidisciplinary research, including atmospheric sciences, soil chemistry and forest ecology. The station consists of a measurement building, a 72 m high mast, a 15 m tall tower and two mini-watersheds. It is equipped with extensive research facilities for measurement of various gases' concentration, various fluxes, meteorological parameters (e.g. temperature, wind speed and direction, relative humidity), solar and terrestrial radiation (e.g. ultraviolet rays), and atmospheric aerosols (e.g. particle size distribution). The measurements for forest ecophysiology and productivity, such as photochemical reflectance, and the measurements for soil and water balance also take place there. A detailed description of the continuous measurements performed at this station can be found in Kulmala et al. (2001a), Hari and Kulmala (2005) and SMEAR website (2017).

2.2 Measured variables

In this study, we used four types of continuous measurement data: gas concentrations, meteorological conditions, radiation variables and aerosol particle concentrations. The gases include nitrogen monoxide (NO) and other oxides (NO_x), ozone (O_3), sulfur dioxide (SO_2), water (H_2O), carbon dioxide (CO_2) and carbon monoxide (CO). Meteorological data include the temperature, humidity, pressure, and wind speed and direction, among others. The gas concentrations and meteorological data measurements are performed at the heights of 4.2, 8.4, 16.8, 33.6, 50.4 and 67.2 m. The radiation variables include UV-A, UV-B, PAR, global, net, reflected global and reflected PAR. These measurements are mostly performed at a radiation tower (18 m). The measured aerosol particle number size distribution ranges were between 3 and 500 nm until December 2004, and after that it has been extended to cover the size range from 3 to 1000 nm. The sampling height was at 2 m until 2015 when the instrument was moved to the tower at 35 m.

Table 1. The name of used atmospheric variables and symbols displayed in the results used in this study.

Atmospheric variables	Symbols
Gas concentrations	
Nitrogen monoxide	NO
Nitrogen oxides	NO_x
Ozone	O_3
Sulfur dioxide	SO_2
Water	H_2O
Carbon dioxide	CO_2
Carbon monoxide	CO
Meteorology	
Rain indicator	SWS
Dew point temperature at 16 m height	T_d
Atmospheric pressure at ground level (180 m above sea level)	Pamb0
Temperature	T_{ave}
Wind speed	WS_{ave}
Wind speed (sonic)	WSU_{ave}
Wind direction	WD_{ave}
Wind direction (sonic)	WDU_{ave}
Relative humidity	$RHURAS_{ave}$
Relative humidity at 16 m height	RHTd
Radiation	
Ultraviolet A	UV_A
Ultraviolet B	UV_B
Diffuse PAR	diffPAR
Diffuse shortwave radiation	diffGLOB
Net radiation	Net
Reflected global radiation	RGlob
Global radiation	Glob
Reflected PAR	RPAR
PAR, total	PAR
Particles concentration	
Aerosol particle number size distribution (3–1000 nm)	3–1000 nm

Table 1 collects all the atmospheric variables used in this study, including the adapted shorthand notation used throughout the current paper together with few details on the measurements. The raw data can be accessed free of charge via the SMEAR website (2017), which also contains more information on the measurements. It should be noted that not all the measured atmospheric variables are included in the current analysis.

2.3 Derived variables

In addition to directly measured variables, there are few derived variables included in this study. The aerosol particle condensation sink determines how rapidly molecules and small particles condense onto preexisting aerosol particles and it is strongly related the shape of the size distribution (Pirjola et al., 1999; Kulmala et al., 2001b). CS is formulated as

$$\mathrm{CS} = 4\pi D \sum_i \beta_{M_i} r_i N_i, \quad (1)$$

where r_i is the radius of a particle for size class i, N_i is the particle concentration in the respective class i, D is the diffusion coefficient of the condensing vapour and β_M is the transitional correction factor, defined in Fuks and Sutugin (1970).

Sulfuric acid (H_2SO_4) concentration is included in the study since it is believed to be one of the key factors in atmospheric aerosol particle formation (Nieminen et al., 2014). Unfortunately, there are no continuous long-term measurements of sulfuric acid concentrations at SMEAR II in Hyytiälä. In order to gauge sulfuric acid, we need to calculate its proxy concentration based on the measured gas concentrations, solar radiation and the measured aerosol particle size distributions acting as CS (Kulmala et al., 2001b). Petäjä et al. (2009) proposed two proxies by using CS and solar radiation in the UV-B range as well as global radiation (Glob). The proxy formulations are given by

$$p_2 = k_2 \cdot \frac{[\mathrm{SO_2}] \cdot \mathrm{UV\text{-}B}}{\mathrm{CS}}, \quad (2)$$

$$p_3 = k_3 \cdot \frac{[\mathrm{SO_2}] \cdot \mathrm{Glob}}{\mathrm{CS}}, \quad (3)$$

where k_2 and k_3 are median values for the scaling factors, which are 9.9×10^{-7} and 2.3×10^{-9} m^2 W^{-1} s, respectively. Here, we include the proxies 2 and 3 (p_2 and p_3) calculated for the years 1996–2014 in our analysis.

Finally, it is essential to have a database of aerosol particle formation days – without such database the correlation analysis between NPF and atmospheric variables cannot be performed. We used a database of the years 1996–2014, generated by the atmospheric scientists at Helsinki University. The database has been created by visual inspection of the continuously measured aerosol size distributions over a size range of 3–1000 nm at the SMEAR II Hyytiälä forest (Dal Maso et al., 2005). The method classifies days into three main groups: event, non-event and undefined days. An event day occurs when there is a growing new mode in the nucleation size range prevailing over several hours, whilst a non-event day takes place when the day is clear of all traces of particle formation. Finally, an undefined day is assumed when it cannot be unambiguously classified as either an event or non-event day. In order to prevent bias in the data, we did not consider the undefined days because this group cannot be unambiguously classified as either an event or non-event day. Undefined days may belong to event or non-event days if further investigation is made. Therefore, the undefined day's group was excluded from our database. Figure 1 shows two examples of the day when NPF and growth (event day) and the day when no particle formation is observed (non-event day) on April 2005 at Hyytiälä station. The x axis displays the 24 h time period whilst y axis denotes the range of particle diameters (from 3 to 1000 nm). The colour indicates the particle concentration level (cm^{-3}).

3 Computational methods: concept and their application

Before the raw data can be fed into an analysis model, they need to be preprocessed first and these steps will be outlined below. After that, the mutual information method will be introduced.

3.1 Data preprocessing

The (raw) data used in this paper range from 1 January 1996 to 31 December 2014, totalling 18 years. The first step in preprocessing is to exclude the undefined days, as the focus is to find the correlation between aerosol particle formation days and atmospheric variables. In order to reduce the amount of irrelevant data, we then eliminate nighttime data points in all atmospheric variables. When the atmospheric photochemistry is most intense (during the daytime), the strongest and long-lasting events of the atmospheric NPF are typically observed (Kulmala and Kerminen, 2008; Nieminen et al., 2014). Due to significant variation in daytime and nighttime in the Hyytiälä forest during a year, it is necessary to use accurate sunrise and sunset times (Duffett-Smith and Zwart, 2011; National Oceanic and Atmospheric Administration, 2017). Since bivariate analysis is performed, between NPF and an atmospheric variable, the time resolution varies for every variable. If a variable is measured every 10 min, it means 10 min time resolution is used.

3.2 Information theory: a brief introduction

Information theory is a mathematical representation of the conditions and parameters affecting the transmission and processing of information (Stone, 2015). It was proposed firstly by Claude E. Shannon in 1948 (Shannon, 1948). Information theory has been applied to a wide range of applications, such as communication (Xie and Kumar, 2004), cryptography (Bruen and Forcinito, 2011) and seismic exploration (Mukerji et al., 2001). Although, information theory has not been used yet to analyse NPF phenomena, this theory has been used in the field of atmospheric sciences, such as acquisition of aerosol size distributions (Preining,

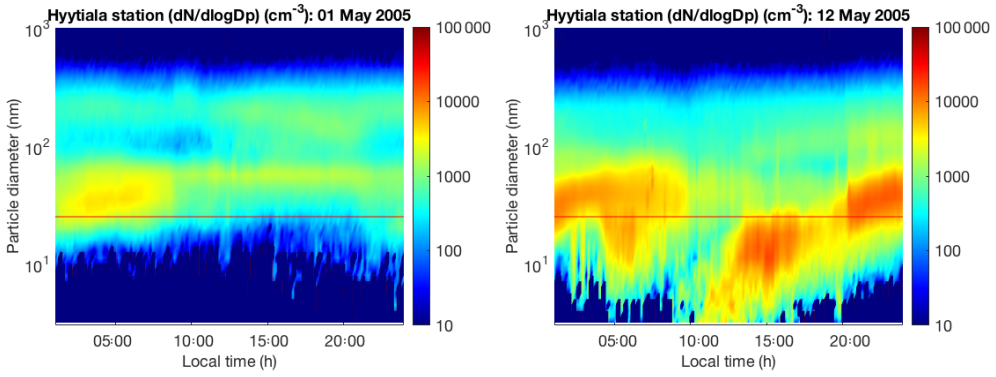

Figure 1. Examples of non-event and event days at Hyytiälä, Finland, in May 2005. A non-event day **(a)** is assumed when the day is clear of all traces of particle formation whilst an event day **(b)** occurs when there is a growing new mode in the nucleation size range prevailing over several hours. Data accessed via Smart-SMEAR (Junninen et al., 2009).

1972), aerosol remote sensing (Li et al., 2009, 2012) and land-precipitation analysis (Brunsell and Young, 2008).

This subsection introduces briefly the basic concepts of information quantities, as well as the definitions and notations of probabilities that will be used throughout the paper. In-depth explanation concerning the principles of information theory can be found for example in MacKay (2003), Cover and Thomas (2012) and Stone (2015).

3.2.1 Entropy

Entropy is a key measure in information theory. It quantifies the amount of uncertainty involved in the value of a random variable. If $\mathbb{X}$ is the set of all data points $\{x_1, \cdots, x_N\}$ that X could take, and $p(x)$ is the probability of some $x \in \mathbb{X}$, then the entropy of X, $H(X)$, is defined as

$$H(X) \equiv -\sum_{x\in\mathbb{X}} p(x)\,\log p(x). \tag{4}$$

Using the concept of information entropy $H(X)$, one can further define two related and useful quantities: the joint and conditional entropies.

Joint entropy measures the amount of uncertainty in two random variables X and Y taken together, and it is defined by

$$H(X,Y) \equiv -\sum_{x\in\mathbb{X},\,y\in\mathbb{Y}} p(x,y)\log p(x,y), \tag{5}$$

where the random variable Y can take values from the set of points $\mathbb{Y} = \{y_1, \cdots, y_N\}$ and $p(x, y)$ is the joint probability of x and y.

Conditional entropy quantifies the amount of uncertainty remaining in the random variable Y when the value of the random variable X is known. This can be defined mathematically by

$$\begin{aligned} H(Y|X) &\equiv -\sum_{x\in\mathbb{X}} p(x)\sum_{y\in\mathbb{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathbb{X},\,y\in\mathbb{Y}} p(x,y)\log\frac{p(x,y)}{p(x)}, \end{aligned} \tag{6}$$

where $p(y|x)$ is the conditional probability of y given x satisfying the chain rule of probability: $p(x, y) = p(y|x)p(x)$. It follows directly from the definition (6) that conditional entropy fulfils the property

$$H(Y|X) = H(X,Y) - H(X), \tag{7}$$

which relates the two-variable conditional and joint entropies with the single-variable information entropy.

3.2.2 Mutual information

The mutual information (MI) of two random variables is a measure of the mutual dependence between these two variables. MI is thus a method for measuring the degree of relatedness between data sets. MI and its relation to joint and conditional entropies is illustrated visually in Fig. 2 with the help of correlated variables X and Y. The left disk (red and orange surface area) shows the entropy $H(X)$, while the right disk (yellow and orange surface area) shows the entropy $H(Y)$. The total surface area covered by the two disks is the joint entropy $H(X, Y)$. The conditional entropy $H(X|Y)$ is the red surface on the left, while the conditional entropy Y given X, $H(Y|X)$, is the yellow surface area on the right. The intersection of the red and yellow disks, the orange surface area in the middle, is the mutual information $I(X; Y)$ between X and Y.

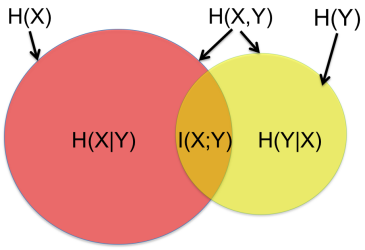

Figure 2. Venn diagram of entropy properties. The area covered by red and orange is the entropy of X, $H(X)$, whilst the area covered by yellow and orange represents the entropy of Y, $H(Y)$. The red area is the conditional entropy of X given by Y, $H(X|Y)$, whereas the yellow area is the conditional entropy of Y given by X, $H(Y|X)$. The area contained by both circles is the joint entropy $H(X,Y)$ and the orange area is the mutual information between X and Y, $I(X;Y)$.

More formally the mutual information of X relative to Y is given as

$$\begin{aligned} I(X;Y) &\equiv H(X,Y) - H(X|Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{8}$$

From the Eq. (8) it is clear that MI is symmetric with respect to the variables X and Y. In terms of probabilities, MI is given by

$$I(X;Y) \equiv \sum_{x\in\mathbb{X}, y\in\mathbb{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}. \tag{9}$$

From the definition (9), one can see that for completely independent and uncorrelated variables, $p(x,y) = p(x)p(y)$, the MI vanishes, as expected. It can be also seen that, in the other extreme where the variables are the same, MI reduces into the corresponding information entropy.

MI has found its use in modern science and technology, for example in search engines (Su et al., 2006), in bioinformatics (Lachmann et al., 2016), in medical imaging (Cassidy et al., 2015) and in feature selection (Peng et al., 2005). Probably at least a part of the MI method's appeal comes from its capability to effectively measure non-linear correlation between data sets (Steuer et al., 2002; Chen et al., 2010). In this aspect MI is superior to the standard Pearson correlation coefficient (PCC) (Pearson, 1895), which is only suitable for measuring linear correlation (Wang et al., 2015). To illustrate this, Fig. 3 shows a comparison between PCC (commonly represented by ρ), the Spearman correlation coefficient (represented by Sp) (Spearman, 1904) and MI using a standard test set of linearly and non-linearly correlated data that is publicly available. The upper row shows six linear data sets, whereas the bottom row plots six non-linear data sets; both rows also contain one uncorrelated data set (the middle one). All methods estimate similar correlation for the linear data sets and correctly detect the uncorrelated data. In the case of the non-linear data, the PCC and Spearman correlation coefficient method simply fail, whereas the MI method is able to measure the correlation in the data.

The MI implementation is straightforward for discrete distributions because the required probabilities for calculating MI can be computed precisely based on counting. However, the MI implementation for continuous distributions may be tricky because the probability distribution function is often unknown. A binning method can be implemented for calculating MI involving continuous distribution. This method makes the data completely discrete by grouping the data points into bins in the continuous variables. Nevertheless, the choice of binning size (i.e. the number of data points per bin) is a non-trivial task, since this choice often leads to different MI result. The binning method does not allow MI calculation between two data sets that have different resolution – this would be a major obstacle in this study. Therefore, in the current investigation we will use the so-called nearest-neighbour method (Kraskov et al., 2004; Ross, 2014). It has been shown to be accurate, insensitive to the choice of model parameter and also computationally relatively fast.

3.3 Mutual information implementation: nearest-neighbour method

This subsection explains the nearest-neighbour MI method adopted from Ross (2014). Suppose x is a discrete variable and y is a continuous variable. The method computes a number I_i for each data point i, based on its nearest neighbours in the continuous variable y. First, using Euclidean distance (or other types of distance metrics), we find the kth closest neighbour to point i among N_{x_i}, where N_{x_i} is the data point whose value of the discrete variable equals x_i. This results in d, that is the distance to this kth neighbour. Next, we count the number of neighbours m_i in the full data set that lie within distance d to point i (including the kth neighbour itself). Based on N_{x_i} and m_i, MI for every data point i can be computed using

$$I_i = \psi(N) - \psi(N_{x_i}) + \psi(k) - \psi(m_i), \tag{10}$$

where N is the number of full data points and k is the user choice for the number of nearest neighbours. The symbol $\psi(.)$ is the digamma function, defined as the logarithmic derivative of the gamma function. This can be expressed as

$$\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\ln\big(\Gamma(z)\big) = \frac{\Gamma'(z)}{\Gamma(z)}, \tag{11}$$

where $\Gamma(.)$ is a gamma function. The detailed explanation about gamma and digamma functions can be found in Abramowitz and Stegun (2012).

After obtaining MI for every point i, in order to estimate the MI from our data set, we average I_i over all data points,

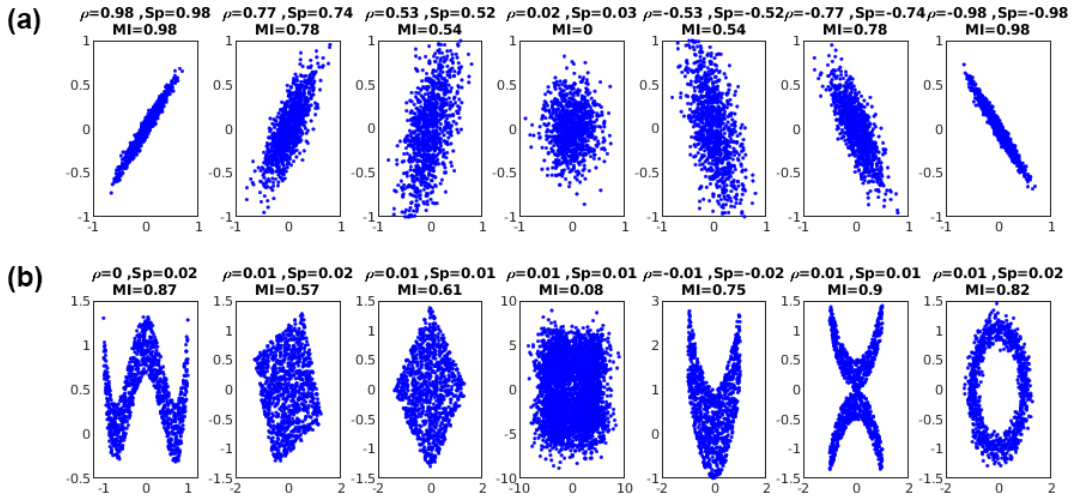

Figure 3. Mutual information (MI) vs. Pearson (ρ) and Spearman (Sp) correlations tested on linear **(a)** and non-linear **(b)** data sets. It can be seen that both methods are able to estimate the correlation for the linear case. On the other hand, the Pearson correlation fails in estimating the correlation for the non-linear data sets, whereas MI is capable of estimating the existence of correlation even in these cases. The so-called nearest-neighbour implementation of the mutual information method is used here (see text).

symbolized by $\langle . \rangle$, to give

$$I(X;Y) = \langle I_i \rangle \tag{12}$$

$$= \psi(N) - \langle \psi(N_x) \rangle + \psi(k) - \langle \psi(m) \rangle, \tag{13}$$

where k is determined by a user. In order to bound the MI estimates within the interval $(-1, 1)$ and make it comparable with the Pearson correlation coefficient (Pearson, 1895), the proposed scaling factor from Numata et al. (2008) is used to give

$$\hat{I}(X;Y) = \mathrm{sign}[I(X;Y)]\sqrt{1-\exp(-2|I(X;Y)|)}, \tag{14}$$

where sign is a signum function and $|.|$ is the absolute value. In this case, the negative values of $\hat{I}(X;Y)$ should not be interpreted as anti-correlations.

Figure 4 illustrates the concept of the nearest-neighbour MI method. This MI implementation is capable of analysing two data sets with different time resolutions. This motivates the adoption of the method in this study, where the time resolution between the measured atmospheric variables and the classification of aerosol particle formation days is not uniform. Hence, the calculation of time-domain features, such as the mean and standard deviation, is not required here. These features naturally compress the data and typically lead to information loss. Panel (a) illustrates the time-series measurement of an atmospheric variable for each day. Every single day can be associated with two classes that are event (E) or non-event days (NE). It can be seen that there are multiple measurements in a day, whereas there are only single event–non-event data available for each day. The distances between the measurement vectors themselves are then calculated as illustrated in panel (b). Here, we take the example of day index 100 (D_{100}). Here, the distance between the measurement vectors at D_{100} from the same class is calculated. In panel (c), the distance vector of D_{100} calculated from the same class (event days) is then ranked in ascending order, shown on the top line. In this particular case, the user choice parameter kth closest neighbour is selected to be 3. So the distance threshold is found at D_{98}. The distance vector from the same class (the red sign) is then projected on the bottom line. The bottom line contains the distance vector of D_{100} calculated from all classes. The dashed line, representing the threshold from point distance D_{100} out to the third neighbour, is drawn until the bottom line. After that, it is found that the number of distance points, which is the third closest neighbour to D_{100} on the top lines, is the seventh closest neighbour on the bottom line ($m = 7$). The above processes point out that the parameter m becomes a crucial factor in the MI estimator, shown in the Eq. (13). This parameter is obtained through the above processes involving the distances calculation between different data resolution. This is advantageous in computing MI between event classification data and atmospheric variable data, which typically vary in different time resolution. In summary, besides its effectiveness in estimating non-linear correlation, the nearest-neighbour MI is also advantageous for the current problem because (1) it is a non-parametric method making no assumptions about the functional form (Gaussian or non-Gaussian) of the statistical distribution underlying the data, (2) there is no need for computationally costly binning to generate histograms, (3) it is computationally fairly light and (4) the model contains only one free model parameter (k) and it is easy to tune.

Prior to demonstrating the result of MI application on the atmospheric data in Sect. 4, the following subsection discusses first how MI is capable of estimating a non-linear relationship, tested on a simulated physics equation.

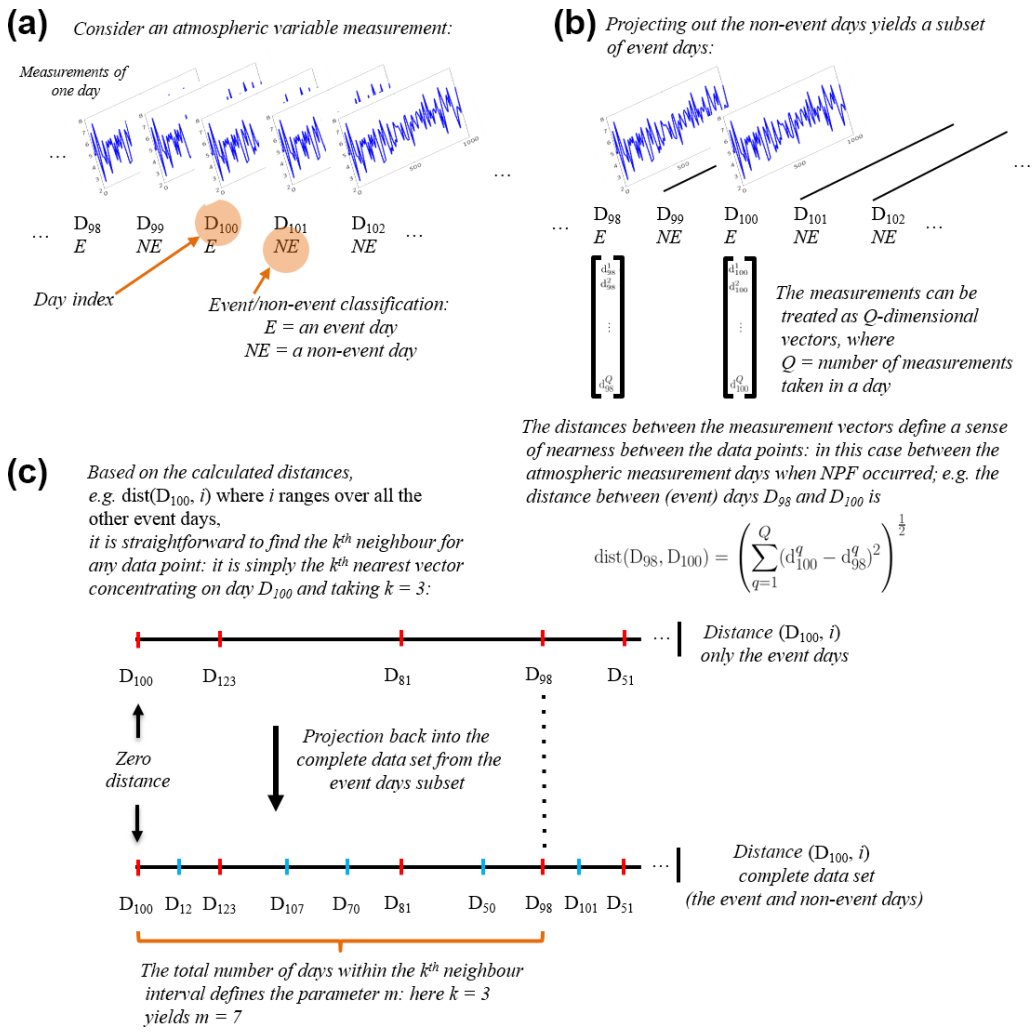

Figure 4. An illustration of the computing process for the mutual information estimator based on the nearest-neighbour method.

3.4 Mutual information: a simulation case study

MI capability in detecting a non-linear relationship between two variables on an artificial benchmark data set is already illustrated in Fig. 3. Before applying nearest-neighbour MI to real atmospheric data, this subsection shows another, more physical case study demonstrating how well MI is able to detect a non-linear relationship between two correlated variables.

We consider the intensity of blackbody radiation. The monochromatic emissive power of a blackbody $F_{\mathrm{B}}(\lambda)$ ($\mathrm{W\,m^{-2}\,\mu m^{-1}}$) is related to temperature T and wavelength λ by (Seinfeld and Pandis, 2016)

$$F_{\mathrm{B}}(\lambda) = \frac{2\pi c^2 h \lambda^{-5}}{e^{ch/k\lambda T} - 1}, \tag{15}$$

where k is the Boltzmann constant ($k = 1.381 \times 10^{-23}\,\mathrm{J\,K^{-1}}$), h is the Planck constant ($6.626 \times 10^{-34}\,\mathrm{Js}$) and c is the speed of light in vacuum ($c = 2.9979 \times 10^{8}\,\mathrm{m\,s^{-1}}$).

The solar spectral irradiance at the top of the Earth's atmosphere at 5777 K is shown in Fig. 5a. If the temperature is varied (randomly between 10 and 10 000 K in this case), the solar spectral irradiance for the same range of wavelengths looks quite different, as is shown in Fig. 5b. The correlation level for both scenarios using PCC (again symbolized by ρ) and the nearest-neighbour MI is also shown. It can be seen that, when the temperature is fixed, the Pearson correlation is still able to detect the correlation between wavelength and solar spectral irradiance, but fails in detecting the relationship between these variables when the data are more messy due to the variation in the temperature. On the other hand, MI is able to detect the correlation between λ and $F_{\mathrm{B}}(\lambda)$ in both cases.

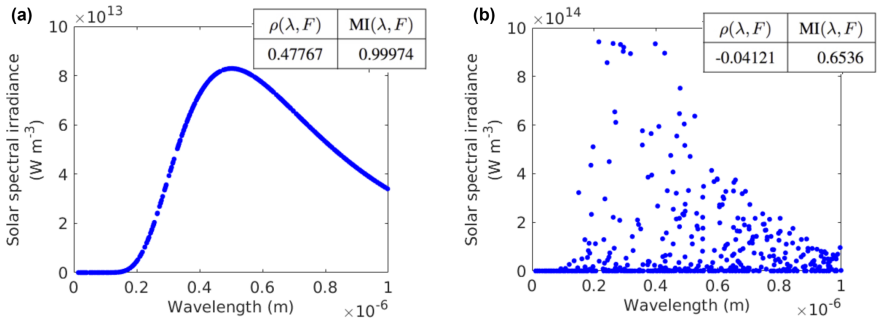

Figure 5. The relationship between solar spectral irradiance and wavelength. The notations ρ and MI represent the Pearson correlation coefficient and mutual information, respectively.

4 Results and discussion

The results section is divided into two subsections. The first part presents the result of MI correlation analysis between atmospheric variables and NPF. The second part then discusses the scatter plot of several relevant atmospheric variables to NPF.

4.1 Correlation analysis between atmospheric variables and NPF

In this study, the atmospheric variables are continuous values while the aerosol formation days classification is discrete. Hence, we implemented the MI based on the nearest-neighbour method for finding the correlation between these two data sets, explained earlier in Sect. 3.3. MI attempts to find the best atmospheric factors/variables which differentiate between event and non-event days. In general, there is no specific level for MI or threshold that indicates a correlation between different variables, which is also similar to the Pearson correlation, where this correlation value gives an only indication of the variables relationship. The value of MI depends on the distribution and the amount data. Unless MI gives a very high value (very close to one) or a very low number (very close to zero), scientists need to make their own judgement about the variable correlation. In this case, similar variables are grouped based on their measurement types (traced gases, radiation, etc.), and their correlation level is ranked. The variables that have the highest MI level indicate that they are more favourable to the NPF process compared to other variables.

Figure 6 presents the correlation results in the form of bar charts, including gases and aerosols (top), meteorology (middle) and radiation (bottom). Several atmospheric variables are measured at different heights, such as gas concentrations and meteorological parameters. In this case, the mean and standard deviation of their MI correlation level were calculated. For those variables, the rectangular bar represents the mean of the MI correlation level, whereas the whisker is its two standard deviations. For the variables which are measured only at one particular height or location, their MI correlation is only represented as the rectangular bar without any whisker.

The top subplot in Fig. 6 shows the MI correlation level between NPF and gas concentrations as well as aerosol (CS). It can be seen that the water concentration (H_2O) has the highest correlation among others. This finding is in agreement with those presented by Boy and Kulmala (2002) and Hyvönen et al. (2005). The reason for the high MI correlation between NPF occurrence and H_2O concentration has so far not been explained. Whether this relation is truly causal or appears because of correlations in diurnal or annual cycles of air masses related to other NPF-related variables remains to be assessed in future studies. The second highest correlation variable in this group is condensation sink. The high correlation with CS can be expected, since CS describes the main sink for vapours participating in NPF and it is also an effective sink for freshly formed new particles. Previous studies have shown that the average value of CS is typically lower on NPF days compared with non-event days (Dal Maso et al., 2007; Asmi et al., 2011b; Dada et al., 2017). Furthermore, this subplot shows that sulfuric acid (H_2SO_4), evaluated using two proxies, correlates well with NPF. It is known that H_2SO_4 is one of the key vapours participating in NPF (Kulmala et al., 2013). The correlation between NPF and H_2SO_4 has been proven through analysis on the data obtained from a number of measurement sites (Kuang et al., 2008; Nieminen et al., 2009; Paasonen et al., 2010; Wang et al., 2011) as well as in laboratory experiments (Almeida et al., 2013).

The MI found that ozone (O_3) and carbon dioxide (CO_2) might be related to the NPF process. The correlations of these variables were also indicated by Hyvönen et al. (2005) via a K-means clustering method. The correlation with O_3 is probably related to the formation of extremely low volatile organic compounds (ELVOCs), which can be initiated by the ozonolysis of monoterpenes (Ehn et al., 2014). ELVOCs are presumed to participate in NPF. The correlation with CO_2, on the other hand, might be related to the coupling between photosynthesis and emission of monoterpenes, as suggested by Kulmala et al. (2014).

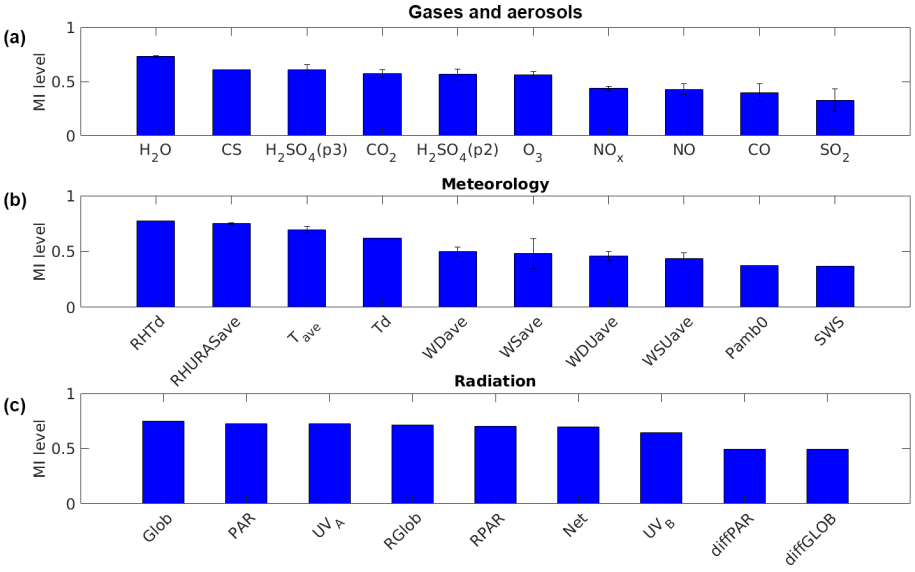

Figure 6. MI correlation level between NPF and a variety of atmospheric variables: gases concentration and aerosols **(a)**, meteorology **(b)** and radiation **(c)**. It can be seen that water concentration (H_2O), condensation sink (CS), sulfuric acid (H_2SO_4), relative humidity (RHTd), average temperature (T_{ave}) and global radiation (Glob) are among the atmospheric variables that have strong correlation to NPF.

On the other hand, the result suggests that sulfur dioxide (SO_2) and nitrogen oxides (NO_x) do not correlate strongly with NPF. The SO_2 observation is inconclusive: its concentration has been found to be higher for NPF event days in some studies (Boy et al., 2008; Young et al., 2013) and lower in others (Wu et al., 2007; Dai et al., 2017). Previously, Boy and Kulmala (2002) already stated that, in the cases of SO_2 and NO_x at this measurement site, there are no significant differences found between event and non-event days.

The middle subplot presents the MI correlation level for all measured meteorological variables. Some variables with the subscript "ave" are averages of meteorological variables measured at different heights. As the top subplot, we calculated the mean and standard deviation of their MI correlation level and display them as a rectangular bar with a whisker. The middle subplot shows that there is a very strong correlation between NPF and relative humidity ($RHURAS_{ave}$ as well as RHTd). A similar result was also reported in Hyvönen et al. (2005). On NPF event days, the average ambient RH is typically lower than non-event days in both clean and polluted environments (Vehkamäki et al., 2004; Hamed et al., 2007; Jun et al., 2014; Qi et al., 2015; Dada et al., 2017). High values of RH tend to have a negative influence on the solar radiation intensity, photochemical reactions and atmospheric lifetime of aerosol precursor vapours (Hamed et al., 2011). Our result points out that the temperature (T_{ave} and T_d) correlates with NPF, as also observed by Boy and Kulmala (2002) and Hyvönen et al. (2005) for this site. The relationship between NPF and temperature may take place due to indirect influences from other factors. For instance, NPF often takes place during the sunny days, when the radiation level and temperature are relatively high. The temperature connection may also occur due to its influence in some chemical reactions leading to NPF. One example might be related emissions of monoterpenes (Tunved et al., 2006; Kiendler-Scharr et al., 2009), which is known as a strong function of temperature (Guenther et al., 1995). However, the temperature is associated with so many atmospheric variables (e.g. boundary layer height, turbulence, radiation, RH and the volatility of the vapours) that the correlation might be caused by several different variables.

In contrast, wind speed (WS_{ave} and WSU_{ave}) and wind direction (WD_{ave} and WDU_{ave}) have little correlation with NPF. Similar results were also reported by Boy and Kulmala (2002). They stated that the small correlation persists due to pollution from the west–southwest (station building and city of Tampere). The correlations between NPF and rain indicator (SWS) as well as the atmospheric pressure (Pamb0) at Hyytiälä were also found to be weak. Several other meteorological variables (not displayed) were excluded from the analysis due to the data scarcity. It is also important to note that on both subplots (top and middle) the whiskers for most bar variables are very short. This means that the MI correlation level for the same variables measured at various heights is similar. The whisker for wind speed (WSU_{ave}) is slightly longer because the measured wind speed varies moderately at different heights.

The bottom subplot shows the MI level of several radiation variables. It can be seen that most radiation variables have a strong relation with NPF. This fact was discussed earlier by Boy and Kulmala (2002), especially on the variable ultraviolet A (UV_A). The high level of correlation in

the global radiation (Glob) was also found by Hyvönen et al. (2005). In all measurement sites, the average solar radiation intensity tends to be higher on NPF event days compared with non-event days (Birmili and Wiedensohler, 2000; Vehkamäki et al., 2004; Hamed et al., 2007; Kristensson et al., 2008; Pierce et al., 2014; Qi et al., 2015; Wonaschütz et al., 2015). Radiation is known as the driving force for atmospheric chemistry, producing low-volatility vapours (e.g. sulfuric acid, ELVOCs) that participate in NPF.

The correlation between concentrations of particles with different sizes from 3 to 1000 nm and NPF is illustrated as a coloured panel in Fig. 7. There are four columns in the x axis. The first three columns represent three periods between years 1996 and 2014, where each period comprises the correlation level for 6 years. The last column is the total correlation level for 18 years. The period division observes if the correlation level for all periods is similar and consistent. The y axis shows the aerosol particle sizes. There are 51 ranges of particles size in the x axis of the coloured panel, but we downsample the 51 particles size ranges to be only 11 sizes for simplification. The colour bar represents the MI correlation level between the specified aerosol particles and NPF. It can be seen that NPF correlates very well with particles in the nucleation mode size range (3–25 nm). This can be expected, since in a relatively clean environment, such as Hyytiälä, NPF is the main source of nucleation mode particles. Clear correlations between particle concentrations and the NPF event occurrence are also detected in the size range from 150 to 550 nm. In this size range, the correlation can be expected, since it is the concentration of these particles that has the largest impact on the condensation sink. Thus, the high concentrations of 150–550 nm particles disfavour NPF and the correlation can be presumed to be negative (see the explanation related to CS in the top panel of Fig. 6).

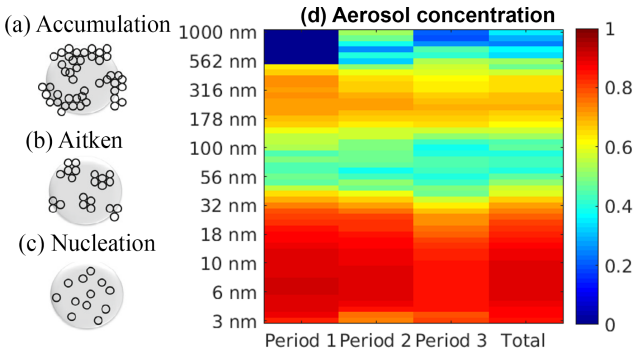

Figure 7. The correlation levels are obtained through the MI method for the particle number size distribution at the SMEAR II station in Hyytiälä. The colour shows the level of correlation. The first three columns represent three periods between years 1996 and 2014, where each period consists of the correlation level for 6 years. The last column indicates the total correlation level for 18 years. Note that the dark blue on period 1 for particles larger than 500 nm is due to unavailable data for that size range.

4.2 Scatter plot analysis

In order to understand in depth the results from the aforementioned MI analysis, a scatter plot matrix was generated, as shown in Fig. 8. The plot involves some of the most important atmospheric variables in the NPF process, according to via the MI analysis made in the previous subsection, including the sulfuric acid concentration (H_2SO_4), average temperature (T_{ave}), relative humidity (RHTd), global radiation (Glob) and condensation sink. The logarithm was applied to the variables H_2SO_4 and CS to ease the scatter plot visualization. The red and blue dots in the plot represent event and non-event days, respectively. Along the diagonal are histogram plots of each column of x. Here, the same data as the above study were used (e.g. 18 years SMEAR II data sets). The undefined days were excluded. Next, the daily mean of all measurements during the daytime was computed and then normalized (between 0 and 1). Finally, for MI comparison, we performed a linear correlation to analyse the relationship between atmospheric variables and event–non-event days. Since the latter is a dichotomous variable (i.e. it contains two categories or discrete), we used the point-biserial correlation coefficient (r_{pb}), which is mathematically equivalent to PCC (Howell, 2012). This correlation coefficient is displayed on each histogram.

First, we focus on the histogram plots located on the subaxes along the diagonal. It can be seen that the event and non-event days are well separated in the cases of Glob and RHTd. These histogram plots demonstrate very well that NPF has a positive ($r_{pb} = 0.639$) and negative ($r_{pb} = -0.707$) correlation with the variables Glob and RHTd, respectively. When the value of global radiation is high, NPF days are likely to occur. On the other hand, non-event days tend to take place when RH is high. The correlations between NPF and these variables were found earlier by MI in the previous subsection. This fact supports the view that the MI method is an effective tool to provide early correlation detection between atmospheric variables and NPF.

The next focus is on the variables H_2SO_4, CS and T_{ave}. The variable H_2SO_4 can still be detected through the linear correlation method ($r_{pb} = 0.403$). The histogram plot of H_2SO_4 shows that NPF event days do not take place when the concentration of H_2SO_4 is very low, whereas the event days usually occur when it is high. However, both event and non-event days may take place if the H_2SO_4 concentration level is medium (i.e. see the intersection between the red and blue histograms). Nevertheless, the scatter plots between Glob, RHTd and H_2SO_4 indicate that these variables are connected in the process of NPF. It is known that the formation of 3 nm particles occurs on the days with strong solar radiation. In other words, to form H_2SO_4 in the atmosphere, high solar radiation is typically required. Likewise, high H_2SO_4 concentration in the atmosphere increases cluster formation and growth rate and hence favours the occurrence of an NPF

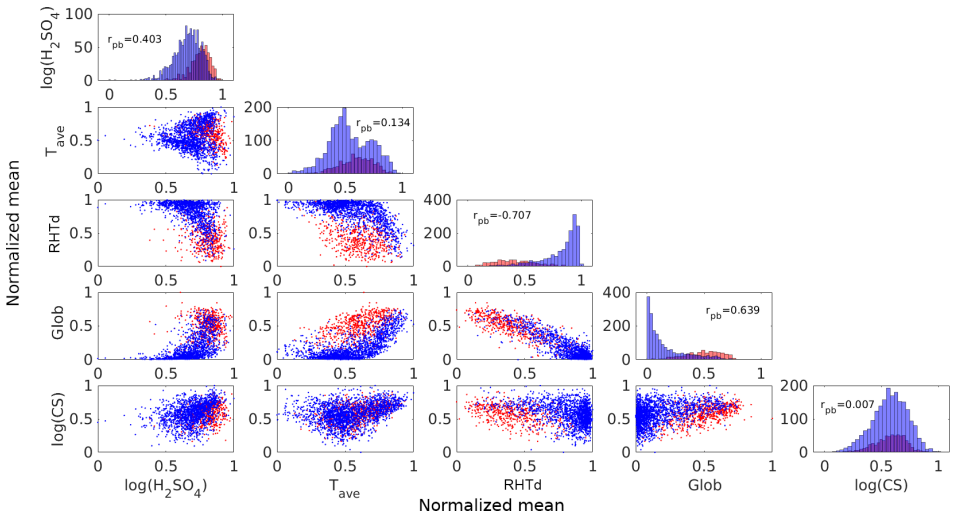

Figure 8. The scatter matrix plot between five selected atmospheric variables and NPF. The red and blue dots represent event and non-event days, respectively. The notation r_{pb} is the point-biserial correlation coefficient.

event (Almeida et al., 2013; Kulmala et al., 2013). On the other hand, when RHTd value is high, the radiation is typically low and therefore the H_2SO_4 concentration also tends to be low.

The above conclusion would be very challenging to make by using a linear correlation analysis for variables T_{ave} ($r_{pb} = 0.134$) and CS ($r_{pb} = 0.007$). These correlation coefficients do not reveal that the variables are related to NPF, which we know from previous literature results. Likewise, by observing the histogram plots, both event and non-event days may take place on any values of T_{ave}, except in very low or very high temperature regimes. This situation is also similar for the case of CS for which event and non-event days are not separable. Since the histogram plots of CS and T_{ave} present the complication in understanding their connection with NPF, their scatter plots should also be analysed. For instance, the event and non-event days seem to be separated on the scatter plots between T_{ave} and Glob as well as RHTd, where the last two variables are known to be correlated with NPF. This may explain how they are connected, but their correlation may be non-linear, since the separation takes place in the middle of the plot. Likewise, the scatter plots between CS and the variables Glob, RHTd and H_2SO_4 show a separation between the event and non-event days. Even though the separation is not perfect, this may still clarify how they are connected.

This subsection demonstrates the analysis complexity by observing the histogram and scatter plots for some variables, such as H_2SO_4, T_{ave} and CS. The intricacy might occur because the relationship among some atmospheric variables and NPF may be complex, non-linear or indirect, in addition to which there might be other variables influencing the process of NPF. CS and temperature are known to impact NPF directly or indirectly, as discussed in Sect. 4.1. A sole investigation through linear correlation analysis, histogram and scatter plots for finding the relationship among atmospheric variables sometimes poses a challenge. This problem explains why MI should be used in the first place for finding early correlation detection between atmospheric variables and their phenomena.

5 Conclusions

This paper extends and complements the analysis of a previous data mining study on atmospheric data, conducted by Hyvönen et al. (2005). Both papers exploit the strengths of data-driven methods, but there are two notable distinctions between this study and the previous investigation. First, our work utilizes 18 years (1996–2014) of atmospheric measurements from the SMEAR II station in Hyytiälä, Finland. This means that the current work deals with 10 more years of data. The utilization of a larger data set is expected to provide more reliable results and thus a more accurate conclusion. Second, instead of using data mining methods based on clustering and classification, this paper promotes the use of MI for identifying the key variables in atmospheric aerosol particle formation. The applied nearest-neighbour MI method is a powerful and computationally light tool capable of finding both linear and non-linear relationships between the measured atmospheric variables and observed NPF events. The method also contains only one free parameter (the number of nearest neighbours, k) and its value does not affect the results significantly (Ross, 2014). Furthermore, the method operates directly on the data and does not require the calculation of characterizing compressed features (i.e. mean, standard deviation) which might potentially lead to a partial information loss.

The MI method reports very similar findings with the previous atmospheric studies. The water content and sulfuric acid concentration are found to be strongly correlated with NPF. Furthermore, the results also suggest that NPF is influenced by temperature, relative humidity, CS and radiation. According to the results from the MI analysis, the measurements taken at different heights have similar correlation with NPF.

As shown in the previous subsection, this method is more powerful than a linear correlation analysis. Therefore, this method should be used in the first place before performing a deeper data analysis method, such as through histogram and scatter plots. This method could act as an early correlation detection for any atmospheric variables.

This work uses the longest available data sets of NPF observations with simultaneously measured ambient variables. As future works, we will seek to investigate the use of the method on different atmospheric data sets. For instance, robust correlation analysis is required for understanding other variables influencing atmospheric process, such as volatile organic compounds (VOCs) and aerosol particles at sizes below 3 nm.

In order to enrich the analysis, the database from other SMEAR stations as well as previous research campaigns should be included. The data may contain more variation because they are measured in different locations. One anticipated obstacle is the scarceness of NPF days classification databases. Although an automatic classification algorithm to create such a database has been called for (Kulmala et al., 2012), currently the event–non-event days are labouriously classified using a manual visualization method (Dal Maso et al., 2005). There has been an attempt to use machine learning for automating aerosol database classification, but the performance has not been completely satisfactory yet (Zaidan et al., 2017). One possibility to enhance the performance of the machine learning classification is to use the correlated atmospheric variables found in this study as additional inputs for such models. Similar concept can also be applied in developing any atmospheric process or proxy. Proxy-dependent variables can be selected by finding the most correlated variables to the interested proxy via MI.

Data availability. Data measured at the SMEAR II station are available on the following web page: https://avaa.tdata.fi/web/smart (last access: 5 August 2018).

Author contributions. MAZ, VH, RR, HJ and ASF designed the study. MAZ and RR developed the methodology. MAZ performed data and statistical analysis. VH, PP, VMK and MK contributed to the interpretation of the data and the results. VH and HJ suggested the use of experimental data from the SMEAR II station. All authors contributed to writing the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work is supported by the European Research Council (ERC) via ATM-GTP (grant number 742206) and the Academy of Finland Centre of Excellence in Atmospheric Sciences (project number 307331).

Edited by: Fangqun Yu
Reviewed by: two anonymous referees

References

Abramowitz, M. and Stegun, I. A.: Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables, Courier Corporation, Washington D.C., 2012.

Almeida, J., Schobesberger, S., Kürten, A., et al.: Molecular understanding of sulphuric acid-amine particle nucleation in the atmosphere, Nature, 502, 359–363, 2013.

Asmi, A., Wiedensohler, A., Laj, P., Fjaeraa, A.-M., Sellegri, K., Birmili, W., Weingartner, E., Baltensperger, U., Zdimal, V., Zikova, N., Putaud, J.-P., Marinoni, A., Tunved, P., Hansson, H.-C., Fiebig, M., Kivekäs, N., Lihavainen, H., Asmi, E., Ulevicius, V., Aalto, P. P., Swietlicki, E., Kristensson, A., Mihalopoulos, N., Kalivitis, N., Kalapov, I., Kiss, G., de Leeuw, G., Henzing, B., Harrison, R. M., Beddows, D., O'Dowd, C., Jennings, S. G., Flentje, H., Weinhold, K., Meinhardt, F., Ries, L., and Kulmala, M.: Number size distributions and seasonality of submicron particles in Europe 2008–2009, Atmos. Chem. Phys., 11, 5505–5538, https://doi.org/10.5194/acp-11-5505-2011, 2011a.

Asmi, E., Kivekäs, N., Kerminen, V.-M., Komppula, M., Hyvärinen, A.-P., Hatakka, J., Viisanen, Y., and Lihavainen, H.: Secondary new particle formation in Northern Finland Pallas site between the years 2000 and 2010, Atmos. Chem. Phys., 11, 12959–12972, https://doi.org/10.5194/acp-11-12959-2011, 2011b.

Birmili, W. and Wiedensohler, A.: New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence, Geophys. Res. Lett., 27, 3325–3328, 2000.

Bonn, B. and Moortgat, G. K.: Sesquiterpene ozonolysis: Origin of atmospheric new particle formation from biogenic hydrocarbons, Geophys. Res. Lett., 30, 1585, https://doi.org/10.1029/2003GL017000, 2003.

Boy, M. and Kulmala, M.: Nucleation events in the continental boundary layer: Influence of physical and meteorological parameters, Atmos. Chem. Phys., 2, 1–16, https://doi.org/10.5194/acp-2-1-2002, 2002.

Boy, M., Karl, T., Turnipseed, A., Mauldin, R. L., Kosciuch, E., Greenberg, J., Rathbone, J., Smith, J., Held, A., Barsanti, K., Wehner, B., Bauer, S., Wiedensohler, A., Bonn, B., Kulmala, M., and Guenther, A.: New particle formation in the Front Range of the Colorado Rocky Mountains, Atmos. Chem. Phys., 8, 1577–1590, https://doi.org/10.5194/acp-8-1577-2008, 2008.

Bruen, A. A. and Forcinito, M. A.: Cryptography, information theory, and error-correction: a handbook for the 21st century, Vol. 68, John Wiley & Sons, Hoboken, New Jersey, 2011.

Brunsell, N. and Young, C.: Land surface response to precipitation events using MODIS and NEXRAD data, Int. J. Remote Sens., 29, 1965–1982, 2008.

Cassidy, B., Rae, C., and Solo, V.: Brain activity: Connectivity, sparsity, and mutual information, IEEE T. Med. Imaging, 34, 846–860, 2015.

Chen, Y. A., Almeida, J. S., Richards, A. J., Müller, P., Carroll, R. J., and Rohrer, B.: A nonparametric approach to detect nonlinear correlation in gene expression, J. Comput. Graph. Stat., 19, 552–568, 2010.

Cover, T. M. and Thomas, J. A.: Elements of information theory, John Wiley & Sons, Hoboken, New Jersey, 2012.

Dada, L., Paasonen, P., Nieminen, T., Buenrostro Mazon, S., Kontkanen, J., Peräkylä, O., Lehtipalo, K., Hussein, T., Petäjä, T., Kerminen, V.-M., Bäck, J., and Kulmala, M.: Long-term analysis of clear-sky new particle formation events and nonevents in Hyytiälä, Atmos. Chem. Phys., 17, 6227–6241, https://doi.org/10.5194/acp-17-6227-2017, 2017.

Dai, L., Wang, H., Zhou, L., An, J., Tang, L., Lu, C., Yan, W., Liu, R., Kong, S., Chen, M., Lee, S., and Yu, H.: Regional and local new particle formation events observed in the Yangtze River Delta region, China, J. Geophys. Res.-Atmos., 122, 2389–2402, 2017.

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P., and Lehtinen, K. E.: Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland, Boreal Environ. Res., 10, 323–336, 2005.

Dal Maso, M., Sogacheva, L., Aalto, P. P., Riipinen, I., Komppula, M., Tunved, P., Korhonen, L., SUUR-USKI, V., Hirsikko, A., Kurtén, T., Kerminen, V.-M., Lihavainen, H., Viisanen, Y., Hansson, H.-C., and Kulmala, M.: Aerosol size distribution measurements at four Nordic field stations: identification, analysis and trajectory analysis of new particle formation bursts, Tellus B, 59, 350–361, 2007.

Duffett-Smith, P. and Zwart, J.: Practical Astronomy with your calculator or spreadsheet, Cambridge University Press, New York, 2011.

Dunne, E. M., Gordon, H., Kürten, A., et al.: Global atmospheric particle formation from CERN CLOUD measurements, Science, 354, 1119–1124, 2016.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., Canagaratna, M., Dal Maso, M., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476–479, 2014.

Fuks, N. A. and Sutugin, A. G.: Highly dispersed aerosols, National Technical Information Service, Springfield, Virginia, 1970.

Guenther, A., Hewitt, C. N., Erickson, D., Fall, R., Geron, C., Graedel, T., Harley, P., Klinger, L., Lerdau, M., McKay, W., Pierce, T., Scholes, B., Steinbrecher, R., Tallamraju, R., Taylor, J., and Zimmerman, P.: A global model of natural volatile organic compound emissions, J. Geophys. Res.-Atmos., 100, 8873–8892, 1995.

Hamed, A., Joutsensaari, J., Mikkonen, S., Sogacheva, L., Dal Maso, M., Kulmala, M., Cavalli, F., Fuzzi, S., Facchini, M. C., Decesari, S., Mircea, M., Lehtinen, K. E. J., and Laaksonen, A.: Nucleation and growth of new particles in Po Valley, Italy, Atmos. Chem. Phys., 7, 355–376, https://doi.org/10.5194/acp-7-355-2007, 2007.

Hamed, A., Korhonen, H., Sihto, S.-L., Joutsensaari, J., Järvinen, H., Petäjä, T., Arnold, F., Nieminen, T., Kulmala, M., Smith, J. N., Lehtinen, K. E. J., and Laaksonen, A.: The role of relative humidity in continental new particle formation, J. Geophys. Res.-Atmos., 116, D03202, https://doi.org/10.1029/2010JD014186, 2011.

Hari, P. and Kulmala, M.: Station for Measuring Ecosystem–Atmosphere Relations (SMEAR II), Boreal Environ. Res., 10, 315–322, 2005.

Howell, D. C.: Statistical methods for psychology, Cengage Learning, Belmont, California, 2012.

Hyvönen, S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T., Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., Pohja, T., Launiainen, S., Hari, P., Mannila, H., and Kulmala, M.: A look at aerosol formation using data mining techniques, Atmos. Chem. Phys., 5, 3345–3356, https://doi.org/10.5194/acp-5-3345-2005, 2005.

Jun, Y.-S., Jeong, C.-H., Sabaliauskas, K., Leaitch, W. R., and Evans, G. J.: A year-long comparison of particle formation events at paired urban and rural locations, Atmos. Pollut. Res., 5, 447–454, 2014.

Junninen, H., Lauri, A., Keronen, P., AaIto, P., HiItunen, V., Hari, P., and Kulmala, M.: Smart-SMEAR: on-line data exploration and visualization tool tor SMEAR stations, Boreal Environ. Res., 14, 447–457, 2009.

Kerminen, V.-M. and Kulmala, M.: Analytical formulae connecting the "real" and the "apparent" nucleation rate and the nuclei number concentration for atmospheric nucleation events, J. Aerosol Sci., 33, 609–622, 2002.

Kerminen, V.-M., Petäjä, T., Manninen, H. E., Paasonen, P., Nieminen, T., Sipilä, M., Junninen, H., Ehn, M., Gagné, S., Laakso, L., Riipinen, I., Vehkamäki, H., Kurten, T., Ortega, I. K., Dal Maso, M., Brus, D., Hyvärinen, A., Lihavainen, H., Leppä, J., Lehtinen, K. E. J., Mirme, A., Mirme, S., Hõrrak, U., Berndt, T., Stratmann, F., Birmili, W., Wiedensohler, A., Metzger, A., Dommen, J., Baltensperger, U., Kiendler-Scharr, A., Mentel, T. F., Wildt, J., Winkler, P. M., Wagner, P. E., Petzold, A., Minikin, A., Plass-Dülmer, C., Pöschl, U., Laaksonen, A., and Kulmala, M.: Atmospheric nucleation: highlights of the EUCAARI project and future directions, Atmos. Chem. Phys., 10, 10829–10848, https://doi.org/10.5194/acp-10-10829-2010, 2010.

Kiendler-Scharr, A., Wildt, J., Dal Maso, M., Hohaus, T., Kleist, E., Mentel, T. F., Tillmann, R., Uerlings, R., Schurr, U., and Wahner, A.: New particle formation in forests inhibited by isoprene emissions, Nature, 461, 381–384, 2009.

Kraskov, A., Stögbauer, H., and Grassberger, P.: Estimating mutual information, Phys. Rev. E, 69, 066138, https://doi.org/10.1103/physreve.83.019903, 2004.

Kristensson, A., Dal Maso, M., Swietlicki, E., Hussein, T., Zhou, J., Kerminen, V.-M., and Kulmala, M.: Characterization of new particle formation events at a background site in Southern Sweden: relation to air mass history, Tellus B, 60, 330–344, 2008.

Kuang, C., McMurry, P., McCormick, A., and Eisele, F.: Dependence of nucleation rates on sulfuric acid vapor concentration in diverse atmospheric locations, J. Geophys. Res.-Atmos., 113, D10209, https://doi.org/10.1029/2007JD009253, 2008.

Kulmala, M.: Build a global Earth observatory, Nature, 553, 21–23, 2018.

Kulmala, M. and Kerminen, V.-M.: On the formation and growth of atmospheric nanoparticles, Atmos. Res., 90, 132–150, 2008.

Kulmala, M., Hämeri, K., Aalto, P., Mäkelä, J., Pirjola, L., Nilsson, E. D., Buzorius, G., Rannik, Ü., Maso, M., Seidl, W., Hoffman, T., Janson, R., Hansson, H.-C., Viisanen, Y., Laaksonen, A., and O'dowd, C. D.: Overview of the international project on biogenic aerosol formation in the boreal forest (BIOFOR), Tellus B, 53, 324–343, 2001a.

Kulmala, M., Maso, M., Mäkelä, J., Pirjola, L., Väkevä, M., Aalto, P., Miikkulainen, P., Hämeri, K., and O'Dowd, C.: On the formation, growth and composition of nucleation mode particles, Tellus B, 53, 479–490, 2001b.

Kulmala, M., Kerminen, V.-M., Anttila, T., Laaksonen, A., and O'Dowd, C. D.: Organic aerosol formation via sulphate cluster activation, J. Geophys. Res.-Atmos., 109, D04205, https://doi.org/10.1029/2003JD003961, 2004.

Kulmala, M., Petäjä, T., Nieminen, T., Sipilä, M., Manninen, H. E., Lehtipalo, K., Dal Maso, M., Aalto, P. P., Junninen, H., Paasonen, P., Riipinen, I., Lehtinen, K. E. J., Laaksonen, A., and Kerminen, V.-M.: Measurement of the nucleation of atmospheric aerosol particles, Nat. Protoc., 7, 1651–1667, 2012.

Kulmala, M., Kontkanen, J., Junninen, H., Lehtipalo, K., Manninen, H. E., Nieminen, T., Petäjä, T., Sipilä, M., Schobesberger, S., Rantala, P., Franchin, A., Jokinen, T., Järvinen, E., Äijälä, M., Kangasluoma, J., Hakala, J., Aalto, P. P., Paasonen, P., Mikkilä, J., Vanhanen, J., Aalto, J., Hakola, H., Makkonen, U., Ruuskanen, T., Mauldin Iii, R. L., Duplissy, J., Vehkamäki, H., Bäck, J., Kortelainen, A., Riipinen, I., Kurtén, T., Johnston, M. V., Smith, J. N., Ehn, M., Mentel, T. F., Lehtinen, K. E. J., Laaksonen, A., Kerminen, V.-M., and Worsnop, D. R.: Direct observations of atmospheric aerosol nucleation, Science, 339, 943–946, 2013.

Kulmala, M., Nieminen, T., Nikandrova, A., Lehtipalo, K., Manninen, H. E., Kajos, M. K., Kolari, P., Lauri, A., Petäjä, T., Krejci, R., Hansson, H.-C., Swietlicki, E., Lindroth, A., Christensen, T. R., Arneth, A., Hari, P., Back, J., Vesala, T., and Kerminen, V.-M.: CO_2-induced terrestrial climate feedback mechanism: From carbon sink to aerosol source and back, Boreal Environ. Res., 19, 122–131, 2014.

Kulmala, M., Kerminen, V.-M., Petäjä, T., Aijun, D., and Wang, L.: Atmospheric Gas-to-Particle Conversion: why NPF events are observed in megacities?, Faraday Discuss., 200, 271–288, 2017.

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A.: ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information, Bioinformatics, 32, 2233–2235, 2016.

Li, Y., Xue, Y., Guang, J., Wang, Y., and Mei, L.: A retrieval algorithm for aerosol optical depth from MODIS multi-spatial scale data based on mutual information, in: Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009, Vol. 5, 489, 2009.

Li, Y., Xue, Y., He, X., and Guang, J.: High-resolution aerosol remote sensing retrieval over urban areas by synergetic use of HJ-1 CCD and MODIS data, Atmos. Environ., 46, 173–180, 2012.

MacKay, D. J.: Information theory, inference and learning algorithms, Cambridge University Press, Cambridge, UK, 2003.

Mikkonen, S., Lehtinen, K. E. J., Hamed, A., Joutsensaari, J., Facchini, M. C., and Laaksonen, A.: Using discriminant analysis as a nucleation event classification method, Atmos. Chem. Phys., 6, 5549–5557, https://doi.org/10.5194/acp-6-5549-2006, 2006.

Mikkonen, S., Korhonen, H., Romakkaniemi, S., Smith, J. N., Joutsensaari, J., Lehtinen, K. E. J., Hamed, A., Breider, T. J., Birmili, W., Spindler, G., Plass-Duelmer, C., Facchini, M. C., and Laaksonen, A.: Meteorological and trace gas factors affecting the number concentration of atmospheric Aitken ($D_p = 50$ nm) particles in the continental boundary layer: parameterization using a multivariate mixed effects model, Geosci. Model Dev., 4, 1–13, https://doi.org/10.5194/gmd-4-1-2011, 2011.

Mukerji, T., Avseth, P., Mavko, G., Takahashi, I., and González, E. F.: Statistical rock physics: Combining rock physics, information theory, and geostatistics to reduce uncertainty in seismic reservoir characterization, The Leading Edge, 20, 313–319, 2001.

National Oceanic and Atmospheric Administration: Sunrise/Sunset Calculator, available at: https://www.esrl.noaa.gov/gmd/grad/solcalc/sunrise.html (last access: 15 December 2017), 2017.

Nieminen, T., Manninen, H., Sihto, S.-L., Yli-Juuti, T., Mauldin, III, R., Petaja, T., Riipinen, I., Kerminen, V.-M., and Kulmala, M.: Connection of sulfuric acid to atmospheric nucleation in boreal forest, Environ. Sci. Technol., 43, 4715–4721, 2009.

Nieminen, T., Asmi, A., Dal Maso, M., Aalto, P. P., Keronen, P., Petaja, T., Kulmala, M., and Kerminen, V.-M.: Trends in atmospheric new-particle formation: 16 years of observations in a boreal-forest environment, Boreal Environ. Res., 19, SS191–SS191, 2014.

Nilsson, E., Paatero, J., and Boy, M.: Effects of air masses and synoptic weather on aerosol formation in the continental boundary layer, Tellus B, 53, 462–478, 2001.

Numata, J., Ebenhöh, O., and Knapp, E.-W.: Measuring correlations in metabolomic networks with mutual information, Genome Inform., 20, 112–122, 2008.

Paasonen, P., Nieminen, T., Asmi, E., Manninen, H. E., Petäjä, T., Plass-Dülmer, C., Flentje, H., Birmili, W., Wiedensohler, A., Hõrrak, U., Metzger, A., Hamed, A., Laaksonen, A., Facchini, M. C., Kerminen, V.-M., and Kulmala, M.: On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation, Atmos. Chem. Phys., 10, 11223–11242, https://doi.org/10.5194/acp-10-11223-2010, 2010.

Pearson, K.: Note on regression and inheritance in the case of two parents, P. R. Soc. London, 58, 240–242, 1895.

Peng, H., Long, F., and Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE T. Pattern Anal., 27, 1226–1238, 2005.

Petäjä, T., Mauldin, III, R. L., Kosciuch, E., McGrath, J., Nieminen, T., Paasonen, P., Boy, M., Adamov, A., Kotiaho, T., and Kulmala, M.: Sulfuric acid and OH concentrations in a boreal forest site, Atmos. Chem. Phys., 9, 7435–7448, https://doi.org/10.5194/acp-9-7435-2009, 2009.

Petäjä, T., Sipilä, M., Paasonen, P., Nieminen, T., Kurtén, T., Ortega, I. K., Stratmann, F., Vehkamäki, H., Berndt, T., and Kulmala, M.: Experimental observation of strongly bound dimers of

sulfuric acid: Application to nucleation in the atmosphere, Phys. Rev. Lett., 106, 228–302, 2011.

Pierce, J. R. and Adams, P. J.: Efficiency of cloud condensation nuclei formation from ultrafine particles, Atmos. Chem. Phys., 7, 1367–1379, https://doi.org/10.5194/acp-7-1367-2007, 2007.

Pierce, J. R., Westervelt, D. M., Atwood, S. A., Barnes, E. A., and Leaitch, W. R.: New-particle formation, growth and climate-relevant particle production in Egbert, Canada: analysis from 1 year of size-distribution observations, Atmos. Chem. Phys., 14, 8647–8663, https://doi.org/10.5194/acp-14-8647-2014, 2014.

Pirjola, L., Kulmala, M., Wilck, M., Bischoff, A., Stratmann, F., and Otto, E.: Effects of aerosol dynamics on the formation of sulphuric acid aerosols and cloud condensation nuclei, J. Aerosol Sci., 30, 1079–1094, 1999.

Preining, O.: Information theory applied to the acquisition of size distributions, J. Aerosol Sci., 3, 289–296, 1972.

Qi, X. M., Ding, A. J., Nie, W., Petäjä, T., Kerminen, V.-M., Herrmann, E., Xie, Y. N., Zheng, L. F., Manninen, H., Aalto, P., Sun, J. N., Xu, Z. N., Chi, X. G., Huang, X., Boy, M., Virkkula, A., Yang, X.-Q., Fu, C. B., and Kulmala, M.: Aerosol size distribution and new particle formation in the western Yangtze River Delta of China: 2 years of measurements at the SORPES station, Atmos. Chem. Phys., 15, 12445–12464, https://doi.org/10.5194/acp-15-12445-2015, 2015.

Ross, B. C.: Mutual information between discrete and continuous data sets, PloS one, 9, e87357, https://doi.org/10.1371/journal.pone.0087357, 2014.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, John Wiley & Sons, Hoboken, New Jersey, 2016.

Shannon, C. E.: A Mathematical Theory of Communication, AT&T Tech. J., 27, 379–423, 1948.

Sipilä, M., Berndt, T., Petäjä, T., Brus, D., Vanhanen, J., Stratmann, F., Patokoski, J., Mauldin, R. L., Hyvärinen, A.-P., Lihavainen, H., and Kulmala, M.: The role of sulfuric acid in atmospheric nucleation, Science, 327, 1243–1246, 2010.

SMEAR website: Station for Measuring Forest Ecosystem Atmosphere Relations, available at: https://www.atm.helsinki.fi/SMEAR/ (last access: 5 August 2018), 2017.

Spearman, C.: The proof and measurement of association between two things, Am. J. Psychol., 15, 72–101, 1904.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J.: The mutual information: detecting and evaluating dependencies between variables, Bioinformatics, 18, S231–S240, 2002.

Stone, J. V.: Information theory: a tutorial introduction, Sebtel Press, Sheffield, UK, 2015.

Su, Q., Xiang, K., Wang, H., Sun, B., and Yu, S.: Using pointwise mutual information to identify implicit features in customer reviews, in: International Conference on Computer Processing of Oriental Languages, 22–30, Springer, Berlin, Heidelberg, 2006.

Tunved, P., Hansson, H.-C., Kerminen, V.-M., Ström, J., Dal Maso, M., Lihavainen, H., Viisanen, Y., Aalto, P., Komppula, M., and Kulmala, M.: High natural aerosol loading over boreal forests, Science, 312, 261–263, 2006.

Vehkamäki, H., Dal Maso, M., Hussein, T., Flanagan, R., Hyvärinen, A., Lauros, J., Merikanto, P., Mönkkönen, M., Pihlatie, K., Salminen, K., Sogacheva, L., Thum, T., Ruuskanen, T. M., Keronen, P., Aalto, P. P., Hari, P., Lehtinen, K. E. J., Rannik, Ü., and Kulmala, M.: Atmospheric particle formation events at Värriö measurement station in Finnish Lapland 1998–2002, Atmos. Chem. Phys., 4, 2015–2023, https://doi.org/10.5194/acp-4-2015-2004, 2004.

Wang, Y., Li, Y., Cao, H., Xiong, M., Shugart, Y. Y., and Jin, L.: Efficient test for nonlinear dependence of two continuous variables, BMC Bioinformatics, 16, 260, https://doi.org/10.1186/s12859-015-0697-7, 2015.

Wang, Z. B., Hu, M., Yue, D. L., Zheng, J., Zhang, R. Y., Wiedensohler, A., Wu, Z. J., Nieminen, T., and Boy, M.: Evaluation on the role of sulfuric acid in the mechanisms of new particle formation for Beijing case, Atmos. Chem. Phys., 11, 12663–12671, https://doi.org/10.5194/acp-11-12663-2011, 2011.

Wang, Z., Wu, Z., Yue, D., Shang, D., Guo, S., Sun, J., Ding, A., Wang, L., Jiang, J., Guo, H., and Gao, J.: New particle formation in China: Current knowledge and further directions, Sci. Total Environ., 577, 258–266, 2017.

Westervelt, D. M., Pierce, J. R., and Adams, P. J.: Analysis of feedbacks between nucleation rate, survival probability and cloud condensation nuclei formation, Atmos. Chem. Phys., 14, 5577–5597, https://doi.org/10.5194/acp-14-5577-2014, 2014.

Wonaschütz, A., Demattio, A., Wagner, R., Burkart, J., Zíková, N., Vodička, P., Ludwig, W., Steiner, G., Schwarz, J., and Hitzenberger, R.: Seasonality of new particle formation in Vienna, Austria–Influence of air mass origin and aerosol chemical composition, Atmos. Environ., 118, 118–126, 2015.

Wu, Z., Hu, M., Liu, S., Wehner, B., Bauer, S., Wiedensohler, A., Petäjä, T., Dal Maso, M., and Kulmala, M.: New particle formation in Beijing, China: Statistical analysis of a 1-year data set, J. Geophys. Res.-Atmos., 112, D09209, https://doi.org/10.1029/2006JD007406, 2007.

Xie, L.-L. and Kumar, P. R.: A network information theory for wireless communication: Scaling laws and optimal operation, IEEE T. Inform. Theory, 50, 748–767, 2004.

Young, L.-H., Lee, S.-H., Kanawade, V. P., Hsiao, T.-C., Lee, Y. L., Hwang, B.-F., Liou, Y.-J., Hsu, H.-T., and Tsai, P.-J.: New particle growth and shrinkage observed in subtropical environments, Atmos. Chem. Phys., 13, 547–564, https://doi.org/10.5194/acp-13-547-2013, 2013.

Zaidan, M. A., Haapasilta, V., Relan, R., Junninen, H., Aalto, P. P., Canova, F. F., Laurson, L., and Foster, A. S.: Neural network classifier on time series features for predicting atmospheric particle formation days, in: The 20th International Conference on Nucleation and Atmospheric Aerosols, edited by: Halonen, R., Nikandrova, A., Kontkanen, J., Enroth, J. A., and Vehkamäki, H., Report Series in Aerosol Science, no. 200, 687–690, 2017.