SMEARcore – Modular data infrastructure for atmospheric measurement stations

Anton Rusanen[1], Kristo Hõrrak[2], Lauri R. Ahonen[1], Tuomo Nieminen[1,3], Pasi P. Aalto[1], Pasi Kolari[1], Markku Kulmala[1,4,5], Tuukka Petäjä[1,5] and Heikki Junninen[2].

[1] Institute for Atmospheric and Earth System Research (INAR) / Physics, Faculty of Science, University of Helsinki, Finland
[2] Institute of Physics, University of Tartu, Estonia
[3] Institute for Atmospheric and Earth System Research (INAR) / Forest Sciences, Faculty of Agriculture and Forestry, University of Helsinki, Finland
[4] Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Sciences and Engineering, Beijing University of Chemical Technology (BUCT), Beijing, China
[5] Joint International Research Laboratory of Atmospheric and Earth System Sciences, School of Atmospheric Sciences, Nanjing University, Nanjing, China

Correspondence to: Anton Rusanen (anton.rusanen@helsinki.fi)

Abstract. We present the SMEARcore data infrastructure framework: a collection of modular programs and processing workflows intended for measurement stations and campaigns as a real-time data analysis and management platform. SMEARcore allows new SMEAR stations (Station for Measuring Earth Surface – Atmosphere Relations) to be built consistently with existing ones and to utilize pre-existing experience in data curation. The structured establishment of new measurements from a data flow point of view allows those stations to directly benefit from our general experience and from further development of visualization and analysis codes. It also establishes robust data pipelines that allow easier diagnosis of problems. We show practical examples of how SMEARcore is utilised at operational measurement stations. This work differs from earlier similar concepts, such as those utilized at stations within ACTRIS (Aerosols, Clouds, and Trace Gases Research Infrastructure) and ICOS (Integrated Carbon Observation System) networks, in three important aspects: Firstly, by keeping all the processing under the control of the data owners. Secondly, by providing tools for making data interoperable in general instead of harmonizing a particular set of instruments and thirdly by being extensible to new instruments. As such it is not meant as a replacement for these infrastructures, but to bring structured data curation to more measurements not covered by them.

1 Introduction

The volume of environmental data doubles faster than every two years (Guo, 2017). Atmospheric composition is continuously monitored with a combination of satellite remote sensing (e.g., Drusch et al. 2012; Beamish et al. 2020) and comprehensive in-situ observations (e.g., Kulmala, 2018; Petäjä et al. 2020). The data is integrated and synthesized in a suite of Earth System models (e.g., Hurrell et al. 2013; Randall et al. 2019). There are ambitions towards digital twin of the whole Earth System

(Bauer et al. 2021), which would enable up-scaling, incorporation of human actions and taking advantage of advances in digital information technology to provide solutions towards sustainable future.

Managing the big data related to the atmosphere is a challenge. Here we place a focus on ground-based in-situ atmospheric observations. In this field, the recent technological advances and particularly a wide use of on-line atmospheric high resolution mass spectrometry allow us to determine concentrations of trace gases and chemical composition of atmospheric aerosol particles with unprecedented accuracy (e.g., Junninen et al. 2010; Yao et al. 2018; Wang et al. 2020). At the same time, there is a constant need for observations at a higher spatial resolution and therefore more stations that provide targeted observations for the region e.g., to tackle specific issues related to air quality or climate change (Kulmala, 2018). More stations and instruments mean also more data. Doing as much processing as possible at the station can aid in the management of the volume of data. Modern atmospheric observations are not single observation points but operated in a network providing harmonized and high-quality data (Laj et al. 2020). To take full advantage of these observations, it is imperative that these systems are well defined, documented and the measurements themselves need to be monitored for measurement instrumentation and hardware malfunctions and anomalies. Measurement systems must also be flexible and facilitate changes in the hardware, personnel, and software as they are inevitable in practice.

In the European scale, topic-specific research infrastructures have been set up to provide harmonized observations, such as Integrated Carbon Observation System (ICOS, Yver-Kwok et al. 2021) and Aerosols, Clouds, and Trace Gases Research Infrastructure (ACTRIS, e.g., Pandolfi et al. 2018). The global perspective is available through World Meteorological Organization's Global Atmospheric Watch (WMO-GAW, Laj et al. 2020). Comprehensive and co-located European infrastructures are endorsed by the SMEAR concept (Hari et al. 2016; Kulmala, 2018). Such stations can be tailored to tackle different grand challenges, such as air quality (e.g., Liu et al. 2020) or climate change (Hari and Kulmala, 2005; Noe et al. 2015).

The various large-scale networks provide mutually different and network-specific standard operation procedures for the stations, ensuring harmonized end user data in their thematic context. The journey from raw data to the data formats provided to end-users is often very labor intensive and done by different people for different instruments. Documenting the steps taken to process the raw data clearly and reproducibly is not simple. The traceability of data deteriorates further when it is used in scientific articles, where reproducibility is a known problem (Buck 2015). Simply put, there can be no reproducibility without proper documentation of what was done. Data analysis, in this paper meaning any process from raw data to products such as end-user data or diagnostics, often lacks such rigor. Developing documented workflows for situations not covered by large-scale network protocols is a problem many stations need to solve.

In this work, we introduce SMEARcore, which aims to answer the problems of data management pertinent both for experimental campaigns and long-term measurement stations such as the SMEAR stations (Hari and Kulmala, 2005). It

provides a set of modular tools for acquiring, transporting, indexing, monitoring, storing, and analyzing data. SMEARcore also provides a consistent interface to access the collected data and enables development of other programs on top of it or embedding the results in, for example, webpages. The interfaces also enable building small networks on top of SMEARcore directly. We illustrate the key features of SMEARcore and show that a station running SMEARcore will provide a streamlined data pipeline from the instruments, measurement computers and databases all the way to the end-user. SMEARcore features functionalities that provide the station managers with real-time updates on the data quality, data collection problems and status of the instruments, measurement computers and accessories. The system indexes this ancillary data, so that it can be accessed for further analysis. This indexing enables the implementation of routine calculations, such as calibrations and visualizations, to be done automatically to aid operators to identify and solve problems with data collection. This standardization of operations and analysis allows us to do science faster, more reliably, and there is a continuous process of supporting metadata and documentation generation for future reference. SMEARcore provides flexible and scalable framework that can be applied at instrument, station or multiple station level.

SMEARcore has 4 main goals: Collect the raw data from disparate sources, monitor and display this process, provide access to this raw data in a common format for further analysis, and do routine analysis to aid operators. We decided to make a modular architecture, which allows us to utilize already existing software solutions whenever possible. It also allows us to program on top of stable interfaces so that the modules are replaceable, and indeed our adjustments to different stations swap the implementation of modules. We hope that in the future this will also allow independent development of data infrastructures based on our interfaces.

2. Workflow

To effectively operate and expand a network of atmospheric stations, the observations need to be supported by coherent data and document management. This is to keep the processes from observational raw data to data products as simple as possible. There seem to be no common tools for getting from measurements into well-structured data that would be widely used in the atmospheric sciences community. Thus, creating a consistent set of processing tools for collecting and processing station data is necessary, and in some subsets of measurements this is already being done (e.g., Mammarella et al., 2016).

Data management is not only about checking that consistent calculations have been made. As with any system, errors can occur, for example: computers crash, power gets interrupted, networks are throttled, reagents run out, somebody forgets to run a script or analyzer inlets foul. Some of these might cause problems for the measurements, some might just temporarily halt data transfers, but in any case, we need to know something unexpected happened. For this to happen in a timely manner, parts of the analysis must be automated and monitored. If we need to wait weeks or months for a responsible person to analyze the data and notice a problem, we cannot intervene when it matters most, and useful data is lost. Same goes for transferring data

out of the measurement computers, monitoring the state of those computers and backups. For these reasons, routine operations should not be a manual process whenever it is possible to automate them.

What does one require from a station scale solution when systemically collecting observational data? The details vary slightly based on what one measures, but in general, the requirements can be summarized into six categories:

1. The ability to collect raw data from several measurement computers.
2. Performing routine analysis that combines several measurements, such as inversions.
3. Storing raw and derived data for an intermittent period for analysis and visualization
4. Displaying this data to the people conducting measurements for quality control
5. Transferring data to long term storage as backup and to vacate local space for new data.
6. Log what files are processed, how, when and present this metadata.

It should be noted that within SMEARcore we assume that the raw data for the analysis is provided by a combination of sensors and instruments, and the associated data acquisition software producing a raw data file. This means we leave making these raw data files up to the data acquisition software.

In practice, a SMEARcore installation is defined by a set of computations, defined as workflows, implementing these steps for a set of instruments and analyses and the backing computational infrastructure. We will now go through some of these workflows to explain what we mean by this concept. It should be noted that the individual workflows are simple, since they are focused on a single purpose. Any complex analysis will utilize the results of previous workflows and the challenge is mostly in coordinating their execution. In practice one usually also includes checks to avoid duplicating work already done in previous workflows, but these are omitted in this paper for clarity. The underlying technical solutions will be described in later chapters.

2.1 Time series data

This is the simplest case of data processing in SMEARcore. The process is visualized in Fig. 1, which contains reading the raw data in and providing the data to visualization and long-term storage in different forms. The input data from different instruments differs mostly in how the instrument-specific raw data format should be interpreted and parsed for visualization, meaning that most instruments can be handled by SMEARcore almost identically. A practical example of such a data process is acquiring total sub-micron aerosol number concentration from a Condensation Particle Counter (e.g., Mordas et al., 2008). Conceptually this dataset is a timeseries of parameters with native averaging from the instrument. The data streams include a timestamp, number concentration, information on the time resolution, and relevant metadata for the instrument and measurement location.

We often create derived datasets from our raw data, for example to do calibrations or calculate new variables. This is accomplished in the workflow in Fig. 1 by having a node that operates on previously collected raw data. In practice plots and collecting several datasets for export in different formats also conform to this pattern, as they are just data products.

2.1.1 Examples of workflows

An example of producing derived datasets is the inversion of the raw datafiles produced by the Differential Mobility Particle Sizer (DMPS; for more details see e.g., Aalto et al., 2001 and Kulmala et al., 2012). The DMPS size selects aerosol particles based on their electrical mobility using differential mobility analyzer (DMA). The concentration of different sized particles is measured by condensation particle counter (CPC). Therefore, the primary data measured by DMPS is particle concentration from CPC at different operating voltages of the DMA, and the voltages need to be converted into a size range. The auxiliary data (various flow rates, pressures and temperatures) which are needed for this inversion are usually stored in the same raw data file (for more details on the inversion process of DMPS raw data, see e.g., Kulmala et al., 2012). This means the required procedure corresponds to the workflow in Fig. 1, with the inversion function handling the processing node. Section 3 explains how the various parts are implemented in SMEARcore.

Another similar, independent, workflow is the processing of flux data from eddy-covariance (EC) measurements. The EC is a technique which utilizes high-frequency measurements of wind and atmospheric variables (e.g., CO_2, H_2O or particle concentrations) for calculating vertical turbulent fluxes between atmosphere and Earth surface. The collected datafiles are 10 Hz raw measurement data, and the EC flux is calculated from the covariance of the fluctuating components of vertical wind and the quantity of interest over some representative time window (typically 30 minutes). The EC data are further processed with the help of auxiliary meteorological data. The creation of derived data involves applying several data processing methods such as detrending, despiking, coordinate rotation, dilution correction and covariance calculations (for detailed descriptions of these methods, see Mammarella et al., 2016). Most atmospheric data processing implemented within SMEARcore can be abstracted to such branching workflows.

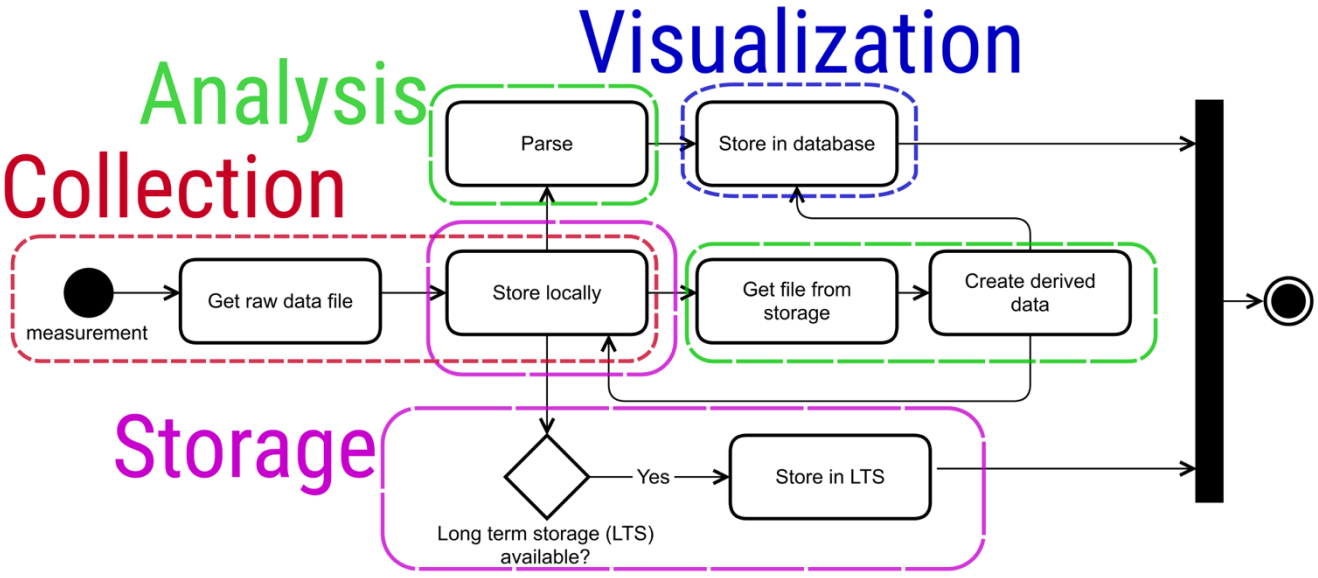


Figure 1. Workflow of processing of a raw file containing time series data and creating derived data products from it. The datafile is collected, parsed and stored, then various further processing can be done to it to create derived data. The different colored hashed boxes indicate which implementation part of SMEARcore is involved in each processing step. The implementation parts are explained in Section 3. The black bordered boxes are steps in the workflow.

2.1.2 Metadata and conventions

Any data files usually need metadata to be interpreted correctly. This is information such as measurement units, calibrations, column names in the data files etc. In our case we also produce metadata about the data processing itself: what files were processed when, how much data was there, what ancillary data was used in the processing and where the files can be located.

One file format used to solve this problem in infrastructures is NASA-AMES used by ACTRIS. There the file metadata is stored within the file itself as extensive header lines. In our case this would lead to extensive duplication of the data in many cases, and it is not appropriate for the metadata about the processing itself. Thus, we store metadata mostly as database tables and link to other files as necessary. Limited sets of metadata can be exported with the files by workflows to produce other formats.

3 Practical execution

In this section we describe the software & hardware used for SMEARcore. Any centralized solution needs at least three things: a server, storage space and a network connecting the computers. The server is responsible for the computing work involved in collecting, indexing, and serving the data to end users. It also provides the platform for any data analysis tasks defined by the workflows. An overview of the system is shown in Fig 2.

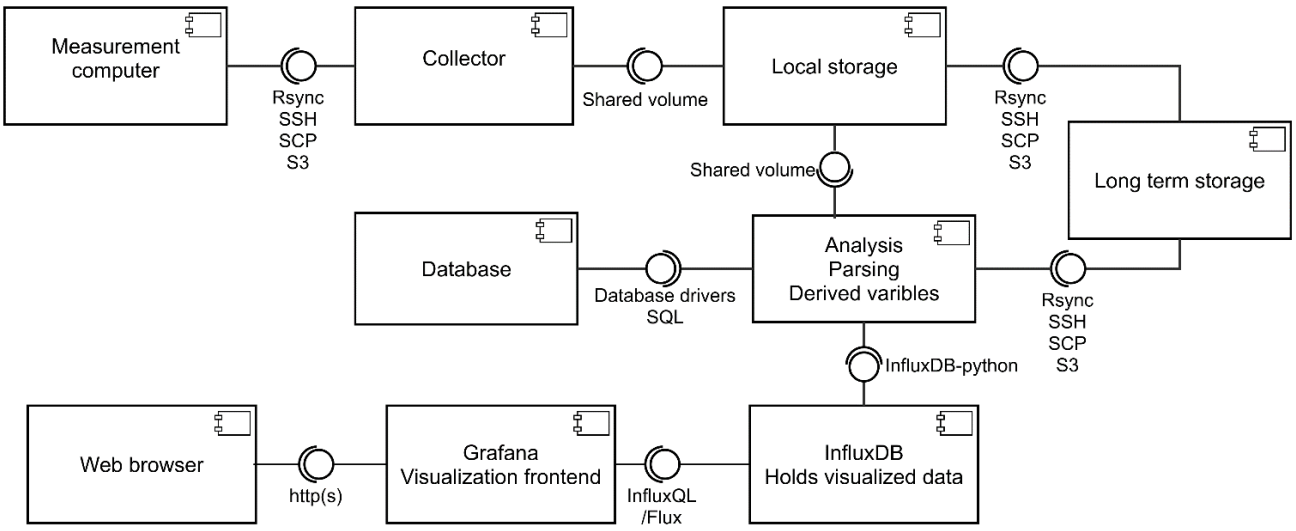


Figure 2. A schematic picture of the different modules and interfaces between them in SMEARcore. The implementation within the box depends on the available hardware on the station, while the interface allows the implementation to change. For example, the analysis does not care where the data is stored if the interface allows retrieving it. The connections with the circle and cup represent the interfaces between the components. The labels in the interface refer to the technologies currently supported in at least one of our SMEARcore installations.

3.1. Software

SMEARcore is built mostly with Python, which is a general-purpose programming language. The choice was based on the permissive license and availability of relevant libraries both for data handling as well as analysis. Python also allows one to call other programs via command line interfaces, which extends our available options by using analysis and processing codes written in other languages.

The entire system is packaged as docker containers, allowing easy installation to a single server via docker compose or on any Kubernetes enabled platform with a series of configuration files. Additional configuration is required to set secrets such as passwords and various routing options as well as selecting the different modular components of the system.

3.1.1 Storage

There are two kinds of storage we need. First is the local storage. This is used to store the files while they are being processed. The other is long-term storage which is used to save the files when they are not in active use. These may reside on the same device or service. Long-term storage may also be completely external such as another service or offline backups. These storages act as the repository of all the measurement data in the system.

We can use multiple backends from local disks to cloud based s3 storage. For local installations failure protection is a desirable feature, so setting up raid for the storage disks and a regular backup schedule is recommended.

3.1.2 Databases

Databases are used for monitoring the service itself, what files have been collected, what workflows have been run on which datafiles and what was produced. They also feed time series data to the plotting software. This means the databases store the status of the station.

We use three different databases: InfluxDB is a timeseries database that integrates with the visualization software and holds the time series data. PostgreSQL or MongoDB holds the information used to coordinate running the workflows. SQLite is used in some versions to hold information about which files have been processed, but this can also be done with the other two databases. The raw data is still stored in the original files and accessed on demand. This is due to a variety of formats, which are not suited for column storage, such as multidimensional fields.

3.1.3 Visualization

For online visualizations we are using Grafana (Grafana 2022). It provides a simple web interface with multiple views that the user can customize. Multiple views allow to get both an overview of the health of the station as well as try to diagnose specific problems by consulting the details from the interface. By default, views are configured to show the status of the measurement computers, as measured by configurable monitoring scripts, and the status of the data collection, as indexed by the SMEARcore database. The interface also allows one to set alarm levels to get a quick notification of the station status.

3.1.4 Data collection

We have currently implemented four different data collection methods for different circumstances:

1. SSH access and rsync based transport. The primary option since it allows the server to control transport, but the measurement computer needs an SSH server and to allow incoming connections.
2. SCP transport from the measurement computer with scripted WinSCP.
3. Shared folders and scripted copying to them on the measurement computer.
4. Storage into an s3 filesystem

Options 2, 3, 4 are similar since there the measurement computer is responsible for the transport and it is difficult to modify without physically going to the computer. They are sometimes necessary due to installation limitations. All forms of transport offer security via passwords or key-based authentication. They also allow us to isolate instruments so that each has their own folders, and they cannot even accidentally overwrite other data. This protects from unintentional data corruption and allows easy management when instruments are changed, or responsible people change.

3.1.5 Analysis

The analysis part of the software runs on top of Apache Airflow (Airflow 2022). This software allows us to define workflows such as calibrations and inversions based on multiple measurement files and run them on a schedule. The software also comes with visualizations of the states of the workflows and possible failures. The workflows themselves are composed of python functions.

There is also an alternative implementation that is done by launching containers directly to do the analysis. In this case each workflow is self-contained.

3.2 Hardware

The server and storage can be co-located and thus far have been in our installations at Estonia (Noe et al. 2015), Beijing (Liu et al. 2020), and Arctic Ocean on board of RV Polarstern (AWI 2017) during the MOSAiC campaign (see e.g., Krumpen, et al., 2021). In these cases, both roles were fulfilled by a Network Attached Storage, NAS, system that supports container virtualization. In general, any computer that supports container virtualization and has enough storage can work as the server.

The hardware parts can also be distributed and run on any cloud provider or external server, allowing functionality without any extra hardware at the station. However, this solution requires a robust network connection from the station. This is the case with our SMEAR III (Järvi et al. 2009) installation, where we use a cloud platform provided by CSC (the Finnish IT Center for Science).

The choice of hardware boils down to how much processing needs to be done and how much data needs to be stored. It is also possible to separate the data collection and do postprocessing on any other platform as is traditionally done. One part of the hardware must be well planned, the local network infrastructure. Since the network serves as the primary way of both transporting the data and representing the status, throttling or disruptions in the network result in degradation of the service. It is possible to run the system without access to the internet, but a local network is still necessary. In practice this usually means setting up routers and wired ethernet connection between the computers.

4. Implementations

4.1 Case study: SMEAR Estonia

The first installation was for the SMEAR station in Järvselja, Estonia (Noe et al. 2015). It uses a centralized server on location and rsync agents installed on the measurement computers so that data can be pulled by the server with data collection method 1 outlined in Section 3.1.4. SMEARcore software containers are run on this server with Docker and are defined using Docker Compose. The data collection and parsing workflows are organized as data acquisition units, DAQs, one pair for each monitored instrument type. The pairs are coordinated using RabbitMQ message queues and a MongoDB database for persistence. SMEARest offers visualization and metadata, such as collected filenames and times, about the collection process through Grafana. There is also direct access to the collected files using sftp and data transfer to off-site storage at University of Tartu high performance computing center.

At the SMEAR Estonia station there are currently 9 instruments which are monitored by SMEARcore at two locations, a main container, and a separate measurement cottage. The instruments measure aerosol properties, meteorological parameters, and background radiation. SMEARcore is integrated with mass spectrometer data analysis software tofTools (Junninen et al. 2010) and allows time-of-flight mass spectrometer data to be processed near-real time (once per hour). This makes it possible to present online processed data as concentrations of chemical groups rather than intensities of single peaks.

The case study is from measurements in SMEAR Estonia station from period June 6th to June 7th, 2021. The figures presented show screenshots from instrument specific real-time dashboards. During this period, we see two daytime new particle formation (NPF) events and a night-time clustering event on 6.6.2021 8pm – 7.6.2021 2am. (Fig. 3 and Fig. 4). The NPF events are seen in the number-size distributions measured by the Neutral cluster and Air Ion Spectrometer (NAIS) as formation of initially 2 nm air ions and their subsequent growth to larger sizes during several hours (Fig. 3). The NAIS data visualization is

done with a custom plugin developed at the SMEAR station. The colormap and the number concentration scale can be changed by the user for their preferred viewing experience. The concentrations of sub-2 nm cluster ions measured by the Cluster Ion Counter (CIC) show diurnal variation, and the concentration of 3 nm ions shows a maximum during NPF events (Fig. 4). During a night-time clustering event the Atmospheric Pressure interface time-of-flight mass spectrometer (CI-APiTOF) observed a simultaneous increase in highly oxidized organic molecules (HOM) dimer concentrations (Fig. 5). Similar night-time clustering events producing small (sub-10 nm) particles which do not grow into larger particles as in daytime NPF events have been observed also at SMEAR II station in Hyytiälä, Finland (Rose et al., 2018).

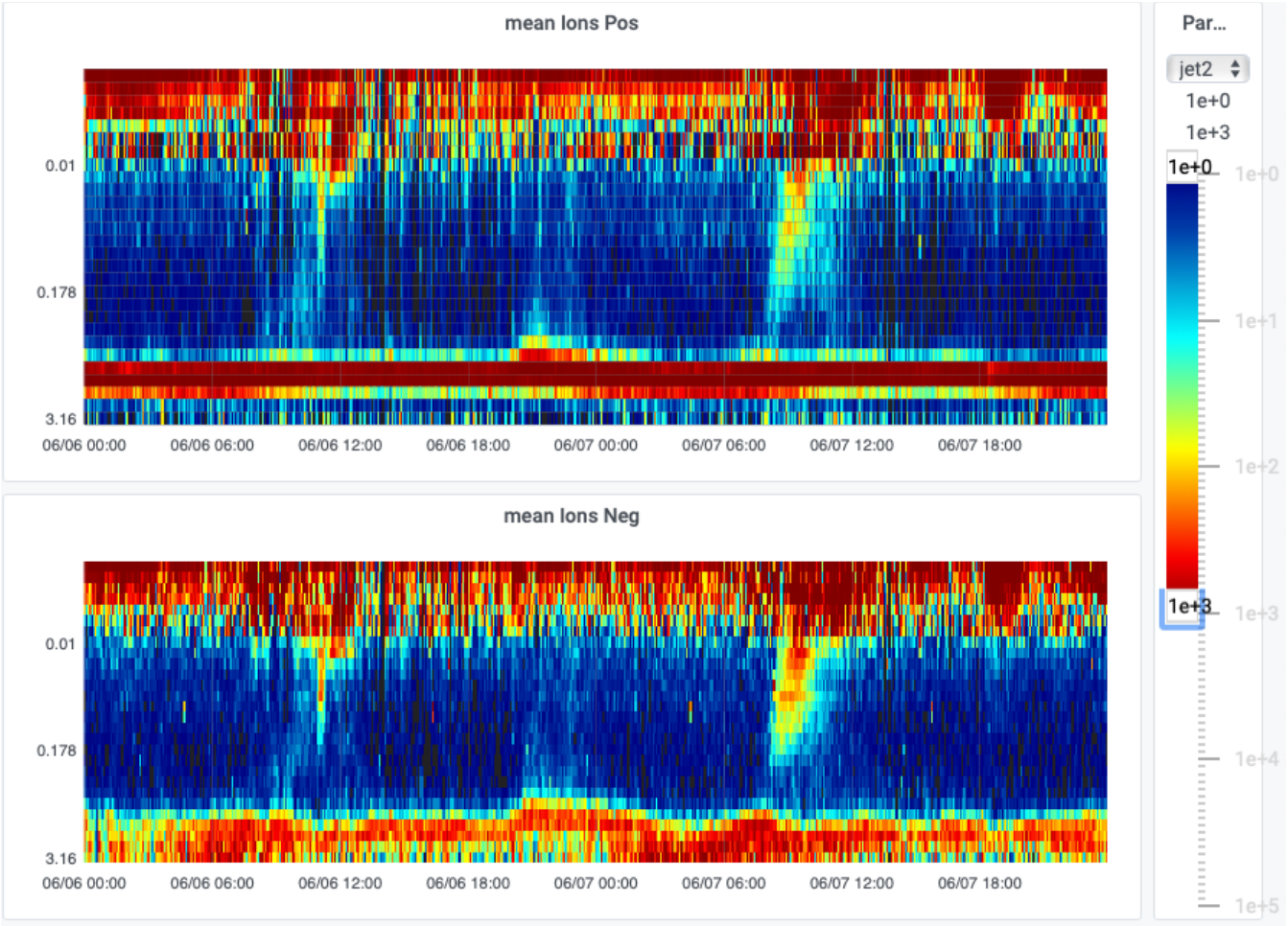
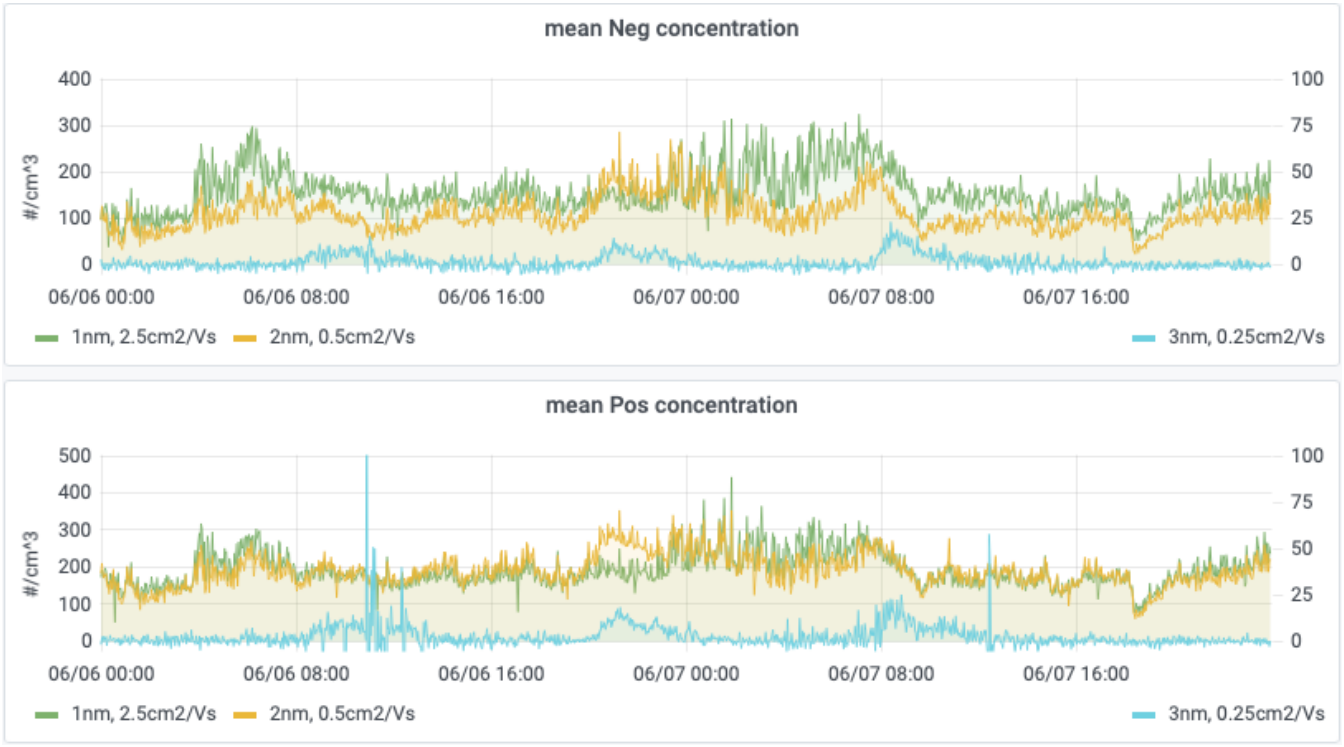


Figure 3. On-line visualization of number size-distribution of positive (top) and negative (bottom) air ions measured with Neutral cluster and Air Ion Spectrometer (NAIS) at Järvselja SMEAR station on 6-7th of June 2021. The y-axis displays in logarithmic scale the ion mobility in units (sV/cm^2). Colorscale indicates the number concentration with units dN/dlog$_{10}$(dp), cm^{-3}.



290 **Figure 4. Concentrations of atmospheric ions with geometric mean diameter of 1 nm (green line) and 2 nm (yellow line) cluster ions, and 3 nm (cyan line) intermediate ions measured at the Järvselja SMEAR station during 6.6.2021 - 7.6.2021 measured by the Cluster Ion Counter (CIC). Negative ions are shown in the top panel and positive ions in the bottom panel.**

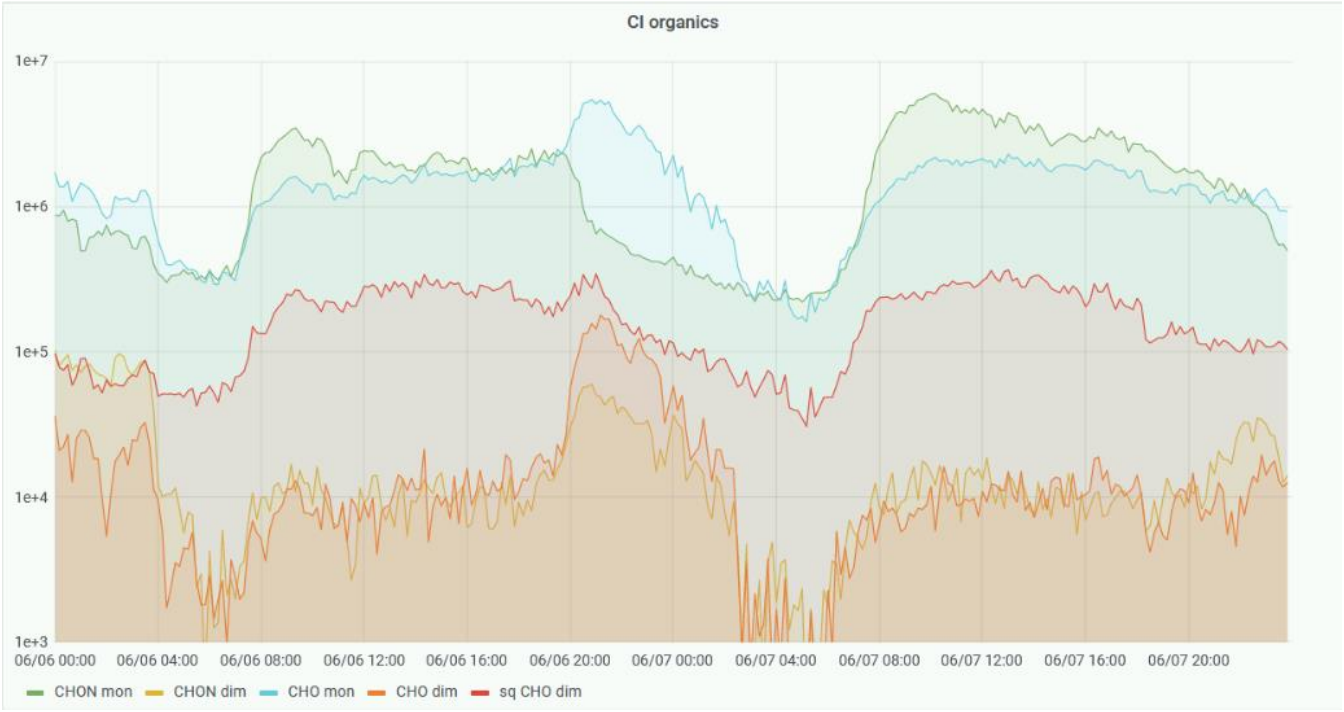


Figure 5. Chemical Ionization Atmospheric Pressure interface Time-Of-Flight (CI-APiTOF) mass spectrometer on-line data of highly oxidized organic molecule (HOM) concentrations measured at the SMEAR Estonia station during 6.6.2021 - 7.6.2021. In the legend capital letters denote chemical elements present in the molecule, "mon" denotes monomers and "dim" dimers, prefix "sq" denotes that HOM are formed from sesquiterpenes. The others are formed from monoterpenes. Both sesquiterpenes and monoterpenes are volatile compounds emitted by vegetation. Inorganic acid concentrations could be presented similarly.

4.1.1 Utilizing Metadata

Figures 6 and 7 show dashboards on the instrument level and station level. An example dashboard from CI-APiTOF mass spectrometer is shown in Figure 6. For the instrument it is important that pressures in different chambers are in the correct range, too low pressure in the first chambers (SSQ and BSQ) indicate clogging of the orifice and mechanical cleaning is required, too high values again indicate malfunction of pumps and pump maintenances is required. In the dashboard monitoring values changes colors from magenta (too low) to green (correct) to red (too high). In addition to current parameter readings also time series of pressure readings are displayed, this helps to identify the reasons for the problem and to see when the problem surfaced. In the case of critical operational parameters an alert is given on the screen and an email is sent to the operator.

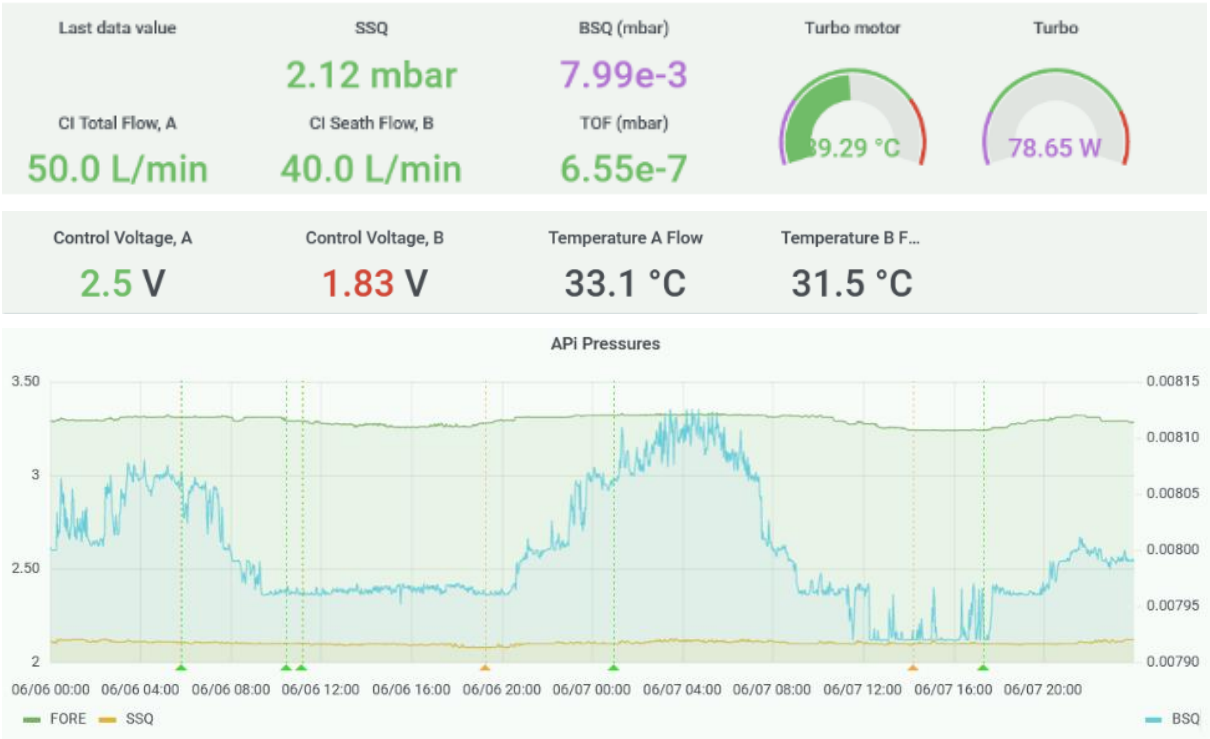


Figure 6. Instrument monitoring parameters for the APi-TOF are plotted on-line. Alerts are triggered if values are out of operation range. The orange and green arrows in the pressure graph indicate that an alert was present.

Various auxiliary measurements can be constantly monitored to ensure the integrity of the measurement devices. Alerts call attention to abnormal readings and can be collected into figures such as Fig. 7. In this dashboard successful file readings are marked with background color, but timing problems due to slow internet or intranet speeds are marked with green color, completely missing files are marked with red. Figure 7 shows current problematic measurements, like GammaTracer and RADOS Cottage and longer lasting problems like with TSI Flow Järvselja, which is not at the station and thus the data is missing. When the instrument is fixed and brought back to the station the status will return to normal.

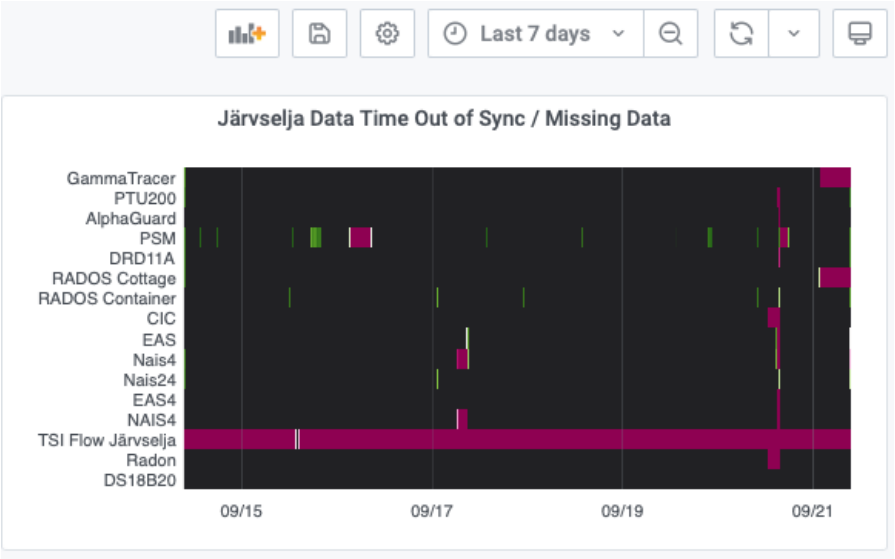


Figure 7. Data syncing statistics from several instruments. Red and green indicate missing and mistimed data respectively. This allows detection of problems at the measurement and network.

In addition to measurement metadata, we collect metadata from the data collection and parsing processes themselves. Last file access times, file sizes, count of parsed columns and how long the parsing took are all things that are monitored. Figure 8 shows an example of how this data can be used to determine the effect of a power cut on collected data. Input Voltage UPS indicates if UPS is being charged, if no input voltage, then system is operated from backup power. Gaps in graphs on 12th and 13th of January are result of such a long power cut, that communication with the station was also interrupted. File size is also smaller due to limited measurement time (Fig. 8).

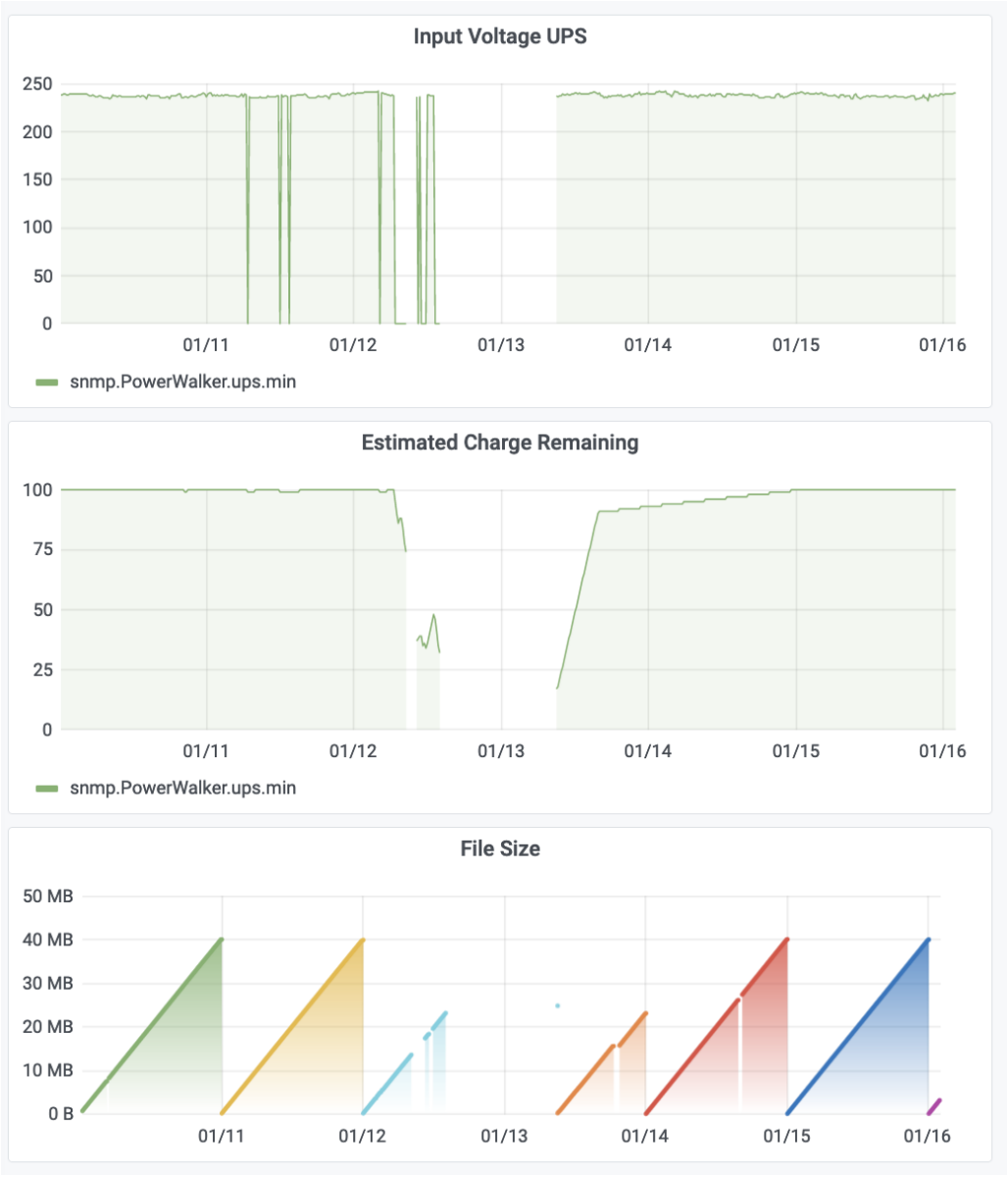


Figure 8. Effects of a long power cut on data collection at the SMEAR Estonia station on 10.-15.1.2021. UPS is fully discharged, and the station experienced network related issues due to no power. The observed data file size is considerably smaller due to missing data, and this is easily diagnosed from the collected metadata.

4.2 Case Study: SMEAR III

Our newest installation is an implementation of parts of the SMEAR III (Järvi et al. 2009) data analysis. The SMEAR III station is located at the University of Helsinki Kumpula campus in Helsinki, Finland. The instruments included are DMPS, basic meteorological instrumentation such as measurements of temperature, wind direction and wind speed, as well as trace gas measurement of ozone, carbon monoxide and nitrogen oxide. The storage backend is s3 in the CSC, IT Center for Science,

cloud service, with a computer simply pushing new files there as they are generated. The data is on a set retention policy, which means that old data is cleared automatically.

Visualization is done in Grafana, and the interface is facing the public internet, allowing the users access from predetermined remote locations. The analysis workflows in Airflow are defined by graphs. Figure 9 shows one such graph from the Airflow user interface. The analysis and visualization components are run in an OpenShift cloud service also at CSC. In this case the main design choice was enabling remote access to users, so the system could not be co-located with the measurements.

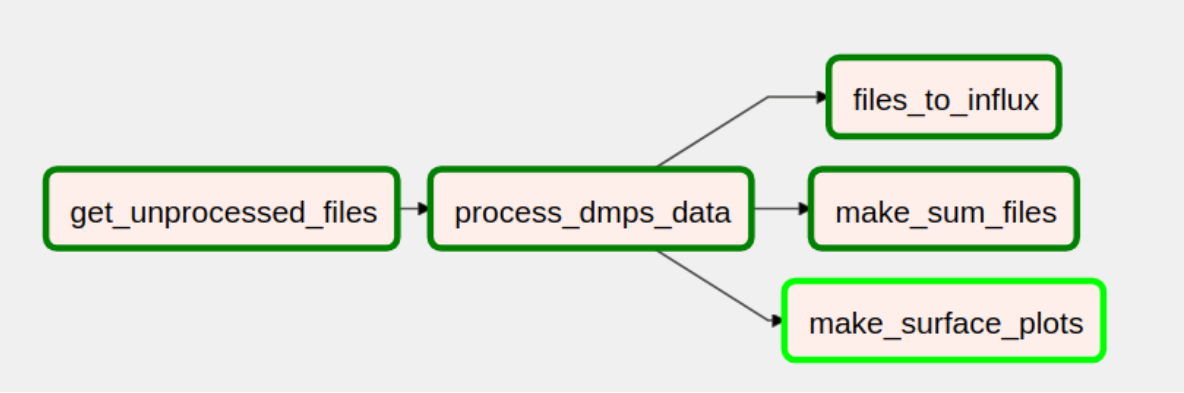


Figure 9. The DMPS processing graph in Airflow. It follows the same structure as the workflow in Figure 2. Influx refers to the database used for storing the processed data. Sum files are the processed file type and surface plots are used for visualization. The arrows represent dependencies, and the last three tasks can be done in parallel. All files are stored in a s3 instance. Colors represent the status of the task. In this case everything but make_surface_plots has completed successfully, while that task is still running.

Figure 10 shows an example visualization of the SMEAR III measurement data. On the morning of 18.2.2021 during 06:00-11:00 a clear surface temperature inversion is evident from the increase in temperature profile from 4 m to 32 m height above ground (Fig. 10a). At the same time, wind speed is also very low (below 1 m s^{-1}), and the wind is from north-to-northeast from the direction of nearby highway a few hundred meters from the measurement site (Fig. 10b, d). The temperature inversion and the low wind speeds lead to inefficient mixing of the air close to ground and accumulation of pollutants emitted from vehicles in the morning traffic and other nearby anthropogenic sources. The particle concentration in the size range 3-820 nm and the concentration of nitrogen oxides (NOx, the sum of NO and NO_2) and carbon monoxide (CO) start increasing between 06:00-08:00, and at the same time the ozone concentration is depleted by more than a factor of 10 to below 1 ppb (Fig. 10c). Similar ground-level ozone depletion episodes have been observed at the Hyytiälä SMEAR II station mostly in autumn and winter connected with low mixing layer, high relative humidity and low solar radiation intensity (Chen et al., 2018). The temperature inversion is strongest around 10:30, when the temperature measured at 32 m height (-10.9°C) is almost 5°C higher compared to temperature at 4 meters (-15.8°C). At this time are observed the highest particle concentrations (above $5 \cdot 10^4$ cm^{-3}) and NOx concentrations (above 200 ppb). The ozone concentration has already started to increase. By 11:30 the temperature inversion has disappeared, the wind speed has increased, and as the result of more efficient vertical and horizontal mixing of the air the measured pollutant concentrations are close to their typical background levels for the Helsinki area. This case study

demonstrates how easy-to-use data visualization tools, which allow efficient comparisons between datasets from multiple instruments, can help in identification of interesting phenomena in the measurements.

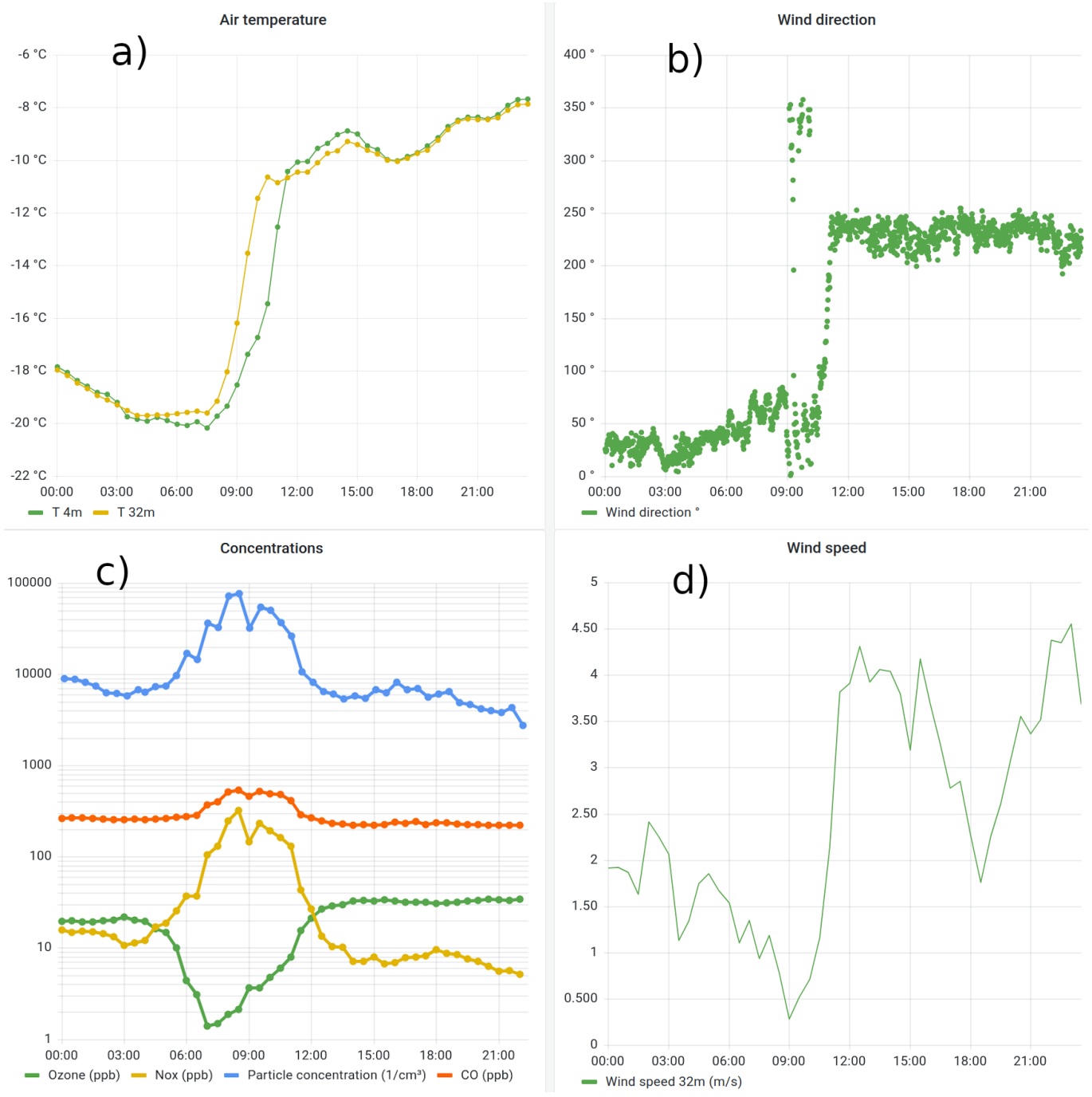


Figure 10. Example of measurements of temperature at heights of 4 m and 32 m (a), wind direction (b), trace gas (ozone, nitrogen oxides and carbon monoxide) and particle concentrations (total concentration in size range 3-820 nm) (c), and wind speed (d) at SMEAR III station in Helsinki, Finland, on 18.2.2021. All the data is visible in a single dashboard.

5. Conclusions

We present a concept for station data management and acquisition using interchangeable components. The concept is in operational use at SMEAR Estonia and has been tested at several campaigns. The components we use are built on popular, well-known open-source projects. This framework is suggested for use at new SMEAR stations and could be useful for larger campaigns as well. Since our system is completely modular, different configurations allow it to be adapted for most common use cases. The system can also easily be extended for more instruments and in the future new technological solutions, as necessary. Compared to centralized solutions such as ICOS or ACTRIS stations, this allows the users to fully control how their data is processed, monitor it in real-time and control how it is transferred outside the station. This makes it useful for measurements not controlled by the centralized solutions. While using our framework does require some technical planning to ensure sufficient hardware resources, we believe the benefits and possibilities of automated data analysis outweigh the costs. We show in two case studies how continuous visualizations of the data and metadata, such as instrument diagnostics and datafile availability, can help quickly spot interesting phenomena and abnormal situations in the measurements.

Since the SMEARcore software allows one to combine multiple data sources, it also provides new opportunities for networking measurement stations together and automatically cross-referencing diverse sources of data in routine operation of the station. This means it is possible to establish smaller networks more easily with the software. An improvement for the management of measurements would be shared storage between stations, where one could check instrument settings or normal operating values at different stations. Another possibility for improving the data usage would be automatically integrating model or satellite data into the analysis or automatically producing the input files for such models, since they can be considered just data products in the SMEARcore framework. In short, automating data processing in the way SMEARcore does also provides opportunities to automate further steps of the scientific process.

6. Author contributions

AR, MK, TP, HJ, PA and PK participated in the initial design of the SMEARcore concept.
MK, TP and HJ participated in funding acquisition, resource acquisition and supervision of the project.
AR, KH, HJ, PA and LA participated in software development and data curation.
AR, KH and HJ investigated the concept by setting up and operating installations.
AR, KH, HJ and TN made the analysis presented in the examples and provided visualizations.
All co-authors participated in the writing and commenting of the manuscript.

7. Competing interests

The authors declare that they have no conflict of interest.

8. Acknowledgements

We acknowledge the following projects: ACCC Flagship funded by the Academy of Finland grant number 337549; Academy professorship funded by the Academy of Finland (grant no. 302958); Academy of Finland projects no. 1325656, 316114 and 325647; "Quantifying carbon sink, CarbonSink+ and their interaction with air quality" INAR project funded by Jane and Aatos Erkko Foundation; European Research Council (ERC) project ATM-GTP Contract No. 742206; and the Arena for the gap analysis of the existing Arctic Science Co-Operations (AASCO) funded by Prince Albert Foundation Contract No 2859.

Technical and scientific staff in Järvselja, Beijing and Hyytiälä stations are acknowledged.

We thank Marjut Kaukolehto for discussions during the planning of SMEARcore.

This work was supported by European Regional Development Fund (MOBTT42), Estonian Research Council (project PRG714) Estonian Environmental Observatory (KKOBS, project 2014-2020.4.01.20-0281), Academy of Finland (grant no. 311932)

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

9. References

Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mäkelä, J. M., Hoell, C., O'dowd, C. D., Hansson, H.-C., Väkevä, M., Koponen, I. K., Buzorius, G., and Kulmala, M.: Physical characterization of aerosol particles during nucleation events, Tellus B: Chemical and Physical Meteorology, 53, 344–358, https://doi.org/10.3402/tellusb.v53i4.17127, 2001.

Airflow: https://airflow.apache.org, last access: 12 January 2022.

Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung.: Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute, Journal of large-scale research facilities, 3, A119, http://dx.doi.org/10.17815/jlsrf-3-163, 2017.

Bauer, P., Stevens B., and Hazeleger. W.: A digital twin of Earth for the green transition, Nature Climate Change 1-4, 2021.

Beamish, A., Raynolds M. K., Epstein, H., Frost, G. V., Macander, M. J., Bergstedt, H., Bartsch, A., Kruse, S., Miles, V., Tanis, C. M., Heim, B., Fuchs, M., Chabrillat, S., Shevtsova, J., Verdonen, M. and Wagner, J.: Recent trends and remaining challenges for optical remote sensing of Arctic tundra vegetation: A review and outlook, Remote Sensing of Environment, https://doi.org/10.1016/j.rse.2020.111872, 2020.

Buck, Stuart.: Solving reproducibility, Science 348: 1403, 2015.

Chen, X., Quéléver, L. L. J., Fung, P. L., Kesti, J., Ri Chen, X., Quéléver, L. L. J., Fung, P. L., Kesti, J., Rissanen, M. P., Bäck, J., Keronen, P., Junninen, H., Petäjä, T., Kerminen, V.-M., and Kulmala, M.: Observations of ozone depletion events in a Finnish boreal forest, Atmospheric Chemistry and Physics, 18, 49–63, https://doi.org/10.5194/acp-18-49-2018, 2018.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P.: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services, Remote Sensing of Environment, 120, 25–36, https://doi.org/https://doi.org/10.1016/j.rse.2011.11.026, 2012.

Grafana: https://grafana.com/grafana, last access: 12 January 2022.

Guo, Huadong: Big Earth data: A new frontier in Earth and information sciences, Big Earth Data 4-20, https://doi.org/10.1080/20964471.2017.1403062 , 2017.

Hari, P., Petäjä, T., Bäck, J., Kerminen, V.-M., Lappalainen, H. K., Vihma, T., Laurila, T., Viisanen, Y., Vesala, T., and Kulmala, M.: Conceptual design of a measurement network of the global change, Atmospheric Chemistry and Physics, 16, 1017-1028, https://doi.org/10.5194/acp-16-1017-2016, 2016.

Hari, P., and Kulmala. M.: Station for measuring ecosystem-atmosphere relations (SMEAR II), Boreal environment research 10: 315-322, 2005.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339-1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Junninen, H., Ehn, M., Petäjä, T., Luosujärvi, L., Kotiaho, T., Kostiainen, R., Rohner, U., Gonin, M., Fuhrer, K., Kulmala, M., and Worsnop, D. R.: A high-resolution mass spectrometer to measure atmospheric ion composition, Atmospheric Measurement Techniques, 3, 1039–1053, https://doi.org/10.5194/amt-3-1039-2010, 2010.

Junninen, H., A. Lauri, P. Keronen, P. Aalto, V. Hiltunen, P. Hari, and M. Kulmala.: Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations, Boreal Environment Research 14: 447-457, 2009.

Järvi, L., Hannuniemi, H., Hussein, T., Junninen, H., Aalto, P. P., Hillamo, R., Mäkelä, T., Keronen, P., Siivola, E., Vesala, T. and Kulmala, M.: The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface-atmosphere interactions in Helsinki, Finland, Boreal Environment Research, 14(SUPPL. A), 86–109, 2009.

Krumpen, T., von Albedyll, L., Goessling, H. F., Hendricks, S., Juhls, B., Spreen, G., Willmes, S., Belter, H. J., Dethloff, K., Haas, C., Kaleschke, L., Katlein, C., Tian-Kunze, X., Ricker, R., Rostosky, P., Rückert, J., Singha, S., and Sokolova, J.: MOSAiC drift expedition from October 2019 to July 2020: sea ice conditions from space and comparison with previous years, The Cryosphere, 15, 3897–3920, https://doi.org/10.5194/tc-15-3897-2021, 2021.

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. Protocols, 7, 1651–1667, https://doi.org/10.1038/nprot.2012.091, 2012.

Kulmala, Markku.: Build a global Earth observatory, Nature 553: 21-23, 2018.

Laj, P., Bigi, A., Rose, C., Andrews, E., Lund Myhre, C., Collaud Coen, M., Lin, Y., Wiedensohler, A., Schulz, M., Ogren, J. A., Fiebig, M., Gliß, J., Mortier, A., Pandolfi, M., Petäja, T., Kim, S.-W., Aas, W., Putaud, J.-P., Mayol-Bracero, O., Keywood, M., Labrador, L., Aalto, P., Ahlberg, E., Alados Arboledas, L., Alastuey, A., Andrade, M., Artíñano, B., Ausmeel, S., Arsov, T., Asmi, E., Backman, J., Baltensperger, U., Bastian, S., Bath, O., Beukes, J. P., Brem, B. T., Bukowiecki, N., Conil, S., Couret, C., Day, D., Dayantolis, W., Degorska, A., Eleftheriadis, K., Fetfatzis, P., Favez, O., Flentje, H., Gini, M. I., Gregorič, A., Gysel-Beer, M., Hallar, A. G., Hand, J., Hoffer, A., Hueglin, C., Hooda, R. K., Hyvärinen, A., Kalapov, I., Kalivitis, N.,

Kasper-Giebl, A., Kim, J. E., Kouvarakis, G., Kranjc, I., Krejci, R., Kulmala, M., Labuschagne, C., Lee, H.-J., Lihavainen, H., Lin, N.-H., Löschau, G., Luoma, K., Marinoni, A., Martins Dos Santos, S., Meinhardt, F., Merkel, M., Metzger, J.-M., Mihalopoulos, N., Nguyen, N. A., Ondracek, J., Pérez, N., Perrone, M. R., Petit, J.-E., Picard, D., Pichon, J.-M., Pont, V., Prats, N., Prenni, A., Reisen, F., Romano, S., Sellegri, K., Sharma, S., Schauer, G., Sheridan, P., Sherman, J. P., Schütze, M., Schwerin, A., Sohmer, R., Sorribas, M., Steinbacher, M., Sun, J., Titos, G., Toczko, B., Tuch, T., Tulet, P., Tunved, P., Vakkari, V., Velarde, F., Velasquez, P., Villani, P., Vratolis, S., Wang, S.-H., Weinhold, K., Weller, R., Yela, M., Yus-Diez, J., Zdimal, V., Zieger, P., and Zikova, N.: A global analysis of climate-relevant aerosol properties retrieved from the network of Global Atmosphere Watch (GAW) near-surface observatories, Atmospheric Measurement Techniques, 13, 4353–4392, https://doi.org/10.5194/amt-13-4353-2020, 2020.

Liu, Y., Yan, C., Feng, Z., Zheng, F., Fan, X., Zhang, Y., Li, C., Zhou, Y., Lin, Z., Guo, Y., Zhang, Y., Ma, L., Zhou, W., Liu, Z., Dada, L., Dällenbach, K., Kontkanen, J., Cai, R., Chan, T., Chu, B., Du, W., Yao, L., Wang, Y., Cai, J., Kangasluoma, J., Kokkonen, T., Kujansuu, J., Rusanen, A., Deng, C., Fu, Y., Yin, R., Li, X., Lu, Y., Liu, Y., Lian, C., Yang, D., Wang, W., Ge, M., Wang, Y., Worsnop, D. R., Junninen, H., He, H., Kerminen, V.-M., Zheng, J., Wang, L., Jiang, J., Petäjä, T., Bianchi, F., and Kulmala, M.: Continuous and comprehensive atmospheric observations in Beijing: a station to understand the complex urban atmospheric environment, Big Earth Data, 4, 295–321, https://doi.org/10.1080/20964471.2020.1798707, 2020.

Mammarella, I., Peltola, O., Nordbo, A., Järvi, L., and Rannik, Ü.: Quantifying the uncertainty of eddy covariance fluxes due to the use of different software packages and combinations of processing steps in two contrasting ecosystems, Atmospheric Measurement Techniques, 9, 4915–4933, https://doi.org/10.5194/amt-9-4915-2016, 2016.

Mordas, G., Manninen, H. E., Petäjä, T., Aalto, P. P., Hämeri, K., and Kulmala, M.: On Operation of the Ultra-Fine Water-Based CPC TSI 3786 and Comparison with Other TSI Models (TSI 3776, TSI 3772, TSI 3025, TSI 3010, TSI 3007), Aerosol Science and Technology, 42, 152–158, https://doi.org/10.1080/02786820701846252, 2008.

Noe, S. M., Niinemets, Ü., Krasnova, A., Krasnov, D., Motallebi, A., Kängsepp, V., Jõgiste, K., Hõrrak, U., Komsaare, K., Mirme, S., Vana, M., Tammet, H., Bäck, J., Vesala, T., Kulmala, M., Petäjä, T., and Kangur, A.: SMEAR Estonia: Perspectives of a large-scale forest ecosystem - atmosphere research infrastructure, Metsanduslikud uurimused, 63, 56–84, 2015.

Pandolfi, M., Alados-Arboledas, L., Alastuey, A., Andrade, M., Angelov, C., Artiñano, B., Backman, J., Baltensperger, U., Bonasoni, P., Bukowiecki, N., Collaud Coen, M., Conil, S., Coz, E., Crenn, V., Dudoitis, V., Ealo, M., Eleftheriadis, K., Favez, O., Fetfatzis, P., Fiebig, M., Flentje, H., Ginot, P., Gysel, M., Henzing, B., Hoffer, A., Holubova Smejkalova, A., Kalapov, I., Kalivitis, N., Kouvarakis, G., Kristensson, A., Kulmala, M., Lihavainen, H., Lunder, C., Luoma, K., Lyamani, H., Marinoni, A., Mihalopoulos, N., Moerman, M., Nicolas, J., O'Dowd, C., Petäjä, T., Petit, J.-E., Pichon, J. M., Prokopciuk, N., Putaud, J.-P., Rodríguez, S., Sciare, J., Sellegri, K., Swietlicki, E., Titos, G., Tuch, T., Tunved, P., Ulevicius, V., Vaishya, A., Vana, M., Virkkula, A., Vratolis, S., Weingartner, E., Wiedensohler, A., and Laj, P.: A European aerosol phenomenology – 6: scattering properties of atmospheric aerosol particles from 28 ACTRIS sites, Atmospheric Chemistry and Physics, 18, 7877–7911, https://doi.org/10.5194/acp-18-7877-2018, 2018.

Petäjä, T., Ganzei, K. S., Lappalainen, H. K., Tabakova, K., Makkonen, R., Räisänen, J., Chalov, S., Kulmala, M., Zilitinkevich, S., Baklanov, P. Y., Shakirov, R. B., Mishina, N. V., Egidarev, E. G., and Kondrat'ev, I. I.: Research agenda for

the Russian Far East and utilization of multi-platform comprehensive environmental observations, International Journal of Digital Earth, 14, 311–337, https://doi.org/10.1080/17538947.2020.1826589, 2021.

Randall, D. A., Bitz, C. M., Danabasoglu, G., Denning, A. S., Gent, P. R., Gettelman, A., Griffies, S. M., Lynch, P., Morrison, H., Pincus, R., and Thuburn, J.: 100 Years of Earth System Model Development, Meteorological Monographs, 59, 12.1 – 12.66, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0018.1, 2018.

Rose, C., Zha, Q., Dada, L., Yan, C., Lehtipalo, K., Junninen, H., Mazon, S. B., Jokinen, T., Sarnela, N., Sipilä, M., Petäjä, T., Kerminen, V.-M., Bianchi, F., and Kulmala, M.: Observations of biogenic ion-induced cluster formation in the atmosphere, Science Advances, 4, eaar5218, https://doi.org/10.1126/sciadv.aar5218, 2018.

Wang, M., Kong, W., Marten, R., He, X.-C., Chen, D., Pfeifer, J., Heitto, A., Kontkanen, J., Dada, L., Kürten, A., Yli-Juuti, T., Manninen, H. E., Amanatidis, S., Amorim, A., Baalbaki, R., Baccarini, A., Bell, D. M., Bertozzi, B., Bräkling, S., Brilke, S., Murillo, L. C., Chiu, R., Chu, B., De Menezes, L.-P., Duplissy, J., Finkenzeller, H., Carracedo, L. G., Granzin, M., Guida, R., Hansel, A., Hofbauer, V., Krechmer, J., Lehtipalo, K., Lamkaddam, H., Lampimäki, M., Lee, C. P., Makhmutov, V., Marie, G., Mathot, S., Mauldin, R. L., Mentler, B., Müller, T., Onnela, A., Partoll, E., Petäjä, T., Philippov, M., Pospisilova, V., Ranjithkumar, A., Rissanen, M., Rörup, B., Scholz, W., Shen, J., Simon, M., Sipilä, M., Steiner, G., Stolzenburg, D., Tham, Y. J., Tomé, A., Wagner, A. C., Wang, D. S., Wang, Y., Weber, S. K., Winkler, P. M., Wlasits, P. J., Wu, Y., Xiao, M., Ye, Q., Zauner-Wieczorek, M., Zhou, X., Volkamer, R., Riipinen, I., Dommen, J., Curtius, J., Baltensperger, U., Kulmala, M., Worsnop, D. R., Kirkby, J., Seinfeld, J. H., El-Haddad, I., Flagan, R. C., and Donahue, N. M.: Rapid growth of new atmospheric particles by nitric acid and ammonia condensation, Nature, 581, 184–189, https://doi.org/10.1038/s41586-020-2270-4, 2020.

Yao, L., Garmash, O., Bianchi, F., Zheng, J., Yan, C., Kontkanen, J., Junninen, H., Mazon, S. B., Ehn, M., Paasonen, P., Sipilä, M., Wang, M., Wang, X., Xiao, S., Chen, H., Lu, Y., Zhang, B., Wang, D., Fu, Q., Geng, F., Li, L., Wang, H., Qiao, L., Yang, X., Chen, J., Kerminen, V.-M., Petäjä, T., Worsnop, D. R., Kulmala, M., and Wang, L.: Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity, Science, 361, 278–281, https://doi.org/10.1126/science.aao4839, 2018.

Yver-Kwok, C., Philippon, C., Bergamaschi, P., Biermann, T., Calzolari, F., Chen, H., Conil, S., Cristofanelli, P., Delmotte, M., Hatakka, J., Heliasz, M., Hermansen, O., Komínková, K., Kubistin, D., Kumps, N., Laurent, O., Laurila, T., Lehner, I., Levula, J., Lindauer, M., Lopez, M., Mammarella, I., Manca, G., Marklund, P., Metzger, J.-M., Mölder, M., Platt, S. M., Ramonet, M., Rivier, L., Scheeren, B., Sha, M. K., Smith, P., Steinbacher, M., Vítková, G., and Wyss, S.: Evaluation and optimization of ICOS atmosphere station data as part of the labeling process, Atmospheric Measurement Techniques, 14, 89–116, https://doi.org/10.5194/amt-14-89-2021, 2021.