Review of Dada et al. “Sources and sinks driving sulphuric acid concentrations in contrasting environments: implications on proxy calculations” by Santtu Mikkonen

The manuscript addresses an important issue on predicting sulphuric acid concentrations when the measurements are not available. Especially finding an applicable proxy for night-time concentrations would be a significant improvement to existing literature. The manuscript introduces different variations of the proposed proxy and they seem to fit nicely on the measurements in selected locations. However, the procedure how the proxy variations were derived and the conditions where the measurements were made need to be described in more detail before the applicability of the proxies can be evaluated and I can recommend the manuscript for publication.

We thank Prof. Santtu Mikkonen for his valuable comments and suggestions, we think that these improve the applicability of the proxy and the overall quality of the study. We provided point-by-point answers in purple. Insertions to the text are in *Italics*. Line numbers refer to the ACPD version of the text.

We thank Santtu again for his constructive comments. In order to address all comments and improve the quality of the manuscript the following developments have been done and their results were added to the manuscript.

To make the following sections straightforward and understandable we start by answering the specific comments 10 and 11 which are relevant to the method section prior to addressing the rest of the comments.

Page 6 lines 251-254: The predictor variables in the proxy contain high measurement uncertainty. Does the fminsearch procedure take that account?

Page 6 lines 254-257: I am happy to see uncertainty estimation for the coefficients made with bootstrap! Though some details on bootstrap procedure should be provided, e.g. how many resamples were drawn?

First of all, the measured data are now divided into independent training and testing data sets. The training sets are used for the derivation of the proxy equations and the testing data sets are used for testing the predictive power of the derived proxies. More details about those data sets are reported in both the main text and in more detail in the supplementary information.

The training sets are measured in Hyytiälä, Agia Marina, Budapest and Beijing. When used for deriving the proxy equation, 10 000 bootstrap resamples were introduced for each data set independently. Bootstrap resampling without disturbance generates extended data from the original data by randomly replacing an existing data point with another one from the same data set, resulting in different combinations of the original data set.

However, the reviewer is right, the fminsearch procedure does not take into account the measurement uncertainty of the predicting variables. Therefore, we included an estimate of error on each of the predictor variable, as well as on H_2SO_4, and included those when generating 10 000 random samples per variable per data point. This was done by scaling the entire time series of a variable by a scalar drawn from a uniform distribution of potential biases of the respective variable (arising for example from uncertainties in calibrations). We did not consider the precision error, since the accuracy error was considerably larger.

Let's take measured sulfuric acid concentration as an example. The measured concentration were accurate within a factor of 2. Therefore, while the temporal behavior of the variable was fairly certain, the entire time series might have been up to a factor too low or up to a factor too high. Therefore, we generated 10 000 concentrations by multiplying the original measured concentration by a uniform random array between the lower and upper bounds, which are 0.5 and 2 in the case of sulfuric acid. The same resampling method was applied for each of the other predictor variables as well as for H_2SO_4 independently, and the 10 000 possible combinations of the disturbed data sets were used to generate 10 000 different *k* value combinations, therefore accounting for the errors in the variables. A median of these 10 000 *k* values was then used to form one equation per location. Additionally, using the testing data sets, we explored whether predicting the concentration varies when we derive the concentration from the median k in the resulting equation, or when we derive it by using the 10 000 *k* values and then taking the median concentration and the difference was negligible. A thorough description of the resampling method is now added to the supplementary information, in addition to the MATLAB code used. The introduction of the uncertainty to the predictor variables and H_2SO_4 widened the range of the 25^{th} and 75^{th} percentiles of the *k* values (Table 1 – ACPD), while narrowing the contribution of each source and sink (Table 2 –ACPD).

The main text Line 251 now reads:

The fitting coefficients were obtained by minimizing the sum of the squared logarithm of the ratio between the proxy values and measured sulphuric acid concentration using the method described by Lagarias et al. (1998), a build-in function fminsearch of MATLAB, giving the optimal values for the coefficients. The data were subject to 10 000 bootstrap resamples when getting each of the k values as a measure of accuracy in terms of bias, variance, confidence intervals, or prediction error (Efron and Tibshirani, 1994). We accounted for the systematic uncertainty in H_2SO_4 and predictor variables. For every bootstrap fit, we assumed both H_2SO_4 and all predictor variables to be affected by independent systematic errors between its lower and upper accuracy limits. More details on the bootstrap resampling method and uncertainty introduction can be found in the supplementary information. The 25^{th} percentile and 75^{th} percentiles of the coefficients are shown for all locations together with the median k values in Table 1. The median k values from the bootstrap resamples were used in the equations for deriving sulphuric acid concentrations at each site.

The complementary section in the SI material now reads:

Bootstrap resampling and sensitivity analyses

When deriving the proxy equation for each site, 10 000 bootstrap resamples were drawn for each data set independently. Bootstrap resampling without disturbance generates extended data from the original data by randomly replacing an existing data point with another one from the same data set, resulting in different combinations of variables from the original data set. We accounted for the systematic uncertainty in H_2SO_4 and predictor variables arising e.g. from calibration uncertainties. For every bootstrap fit, we assumed both H_2SO_4 and all predictor variables to be affected by independent systematic errors between the upper and lower bound of their independent uncertainty ranges. Since the uncertainty related to the measurement accuracy was much larger than the precision of the measurement, we only accounted for the uncertainty arising from accuracy. In practice, we scaled the entire time series of each variable by a random set of numbers drawn from a uniform distribution of possible measurement biases.

Accordingly, a factor of 2 uncertainty was introduced in the sulphuric acid concentration, a 20% uncertainty in the condensation sink measurement, and a 10% in each trace gas concentration and global radiation. In the case of sulphuric acid concentrations, which have a factor of 2 uncertainty, the actual concentration of sulphuric acid at a certain point in time could be anywhere between a factor of 2 lower and a factor of 2 higher. Therefore, for each sulphuric acid measurement, we generated 10 000 concentrations by multiplying the original measured concentration by a uniform random array between the lower and upper bounds, which are 0.5 and 2 in the case of sulphuric acid. The same resampling method ws applied for each other predictor variable independently, and the 10 000 possible combinations of the disturbed data sets were used to generate the fit and to derive the sulphuric acid proxy equation per site. A median of these 10 000 k value combinations which account for the error on the predictor variables was then used to form one equation per location. The MATLAB code used to generate the boot resamples is shown in Code 1.

Major Comments

1. The proxies for individual campaigns were derived from the same data they are predicting, these proxies need to be verified on independent data before they can be generalized even on different conditions in the same sites.
2. In addition, the data were collected from short periods, except for Hyytiälä, and it would be helpful if there would be some discussion on how representative the measurements are compared to annual level or long term seasonal averages of all variables in the sites. Bootstrap resampling is good method in the case where not so much comparable data are available but it is not enough for constructing a generalizable tool if the measurements are not representative.
3. Derivation of night-time proxies in Hyytiälä should be revisited. I would suggest calculating separate proxies for dark time without global radiation included, or similarly than in China, as the chemistry is different during the dark hours. The manuscript suggests that the night-time formation of sulphuric acid is mostly driven by Criegee intermediates and thus the coefficient k2 in China was seen to be significantly higher than for daytime and that might be the case also in Hyytiälä.

1. We explored the predictive power of our proxy by testing it on independent data sets.

Each of the proxies of the boreal forest environment, rural background and mega city are tested for predictive power on independent data sets using extended data sets from the same location or using measurements from locations with similar characteristics (CS, trace gas concentrations – reference to Figure 10 in ACPD version). However, unfortunately our group has not performed any recent measurements in an urban location similar to the one in Budapest with a similar instrument or calibration, therefore for this specific site, we rely on bootstrap resampling only for accounting for variability in the predictor variables (Figures R1 – R15.

Overall, the modelled sulphuric acid concentrations correlated well (R = 0.7- Boreal; R = 0.45 – Rural and R = 0.83 – Megacity) with the measured sulphuric concentrations with a slope of ~1 for the testing data set except for the rural site, which could be attributed to the missing alkene source term resulting from the absence of alkene measurement in the Agia Marina data set. Additionally, we found that for all of the three testing data sets, the difference between the measured and modelled sulphuric acid concentrations was less than the error on the predication model itself for almost 70% of the data points. Note that the model prediction error was estimated as the interquartile range of the modelled H_2SO_4 concentration of a single point in time arising from the 10 000 different combinations in *k* values (Figures R2, R6, R9 and R13).

1.1. Boreal environment:

The training data set used to develop the proxy equation was from August 18, 2016 to December 31, 2016 and from March 8, 2018 to February 28, 2019. For testing the predictive power of the proxy, we used an independent testing data set from January 1, 2017 to June 5, 2017 from the same location.

Hyytiälä proxy Equation 9:.

$$[H_2SO_4]_{boreal} = -\frac{CS}{2\,x\,(4.2\,x\,10^{-9})} + \left[\left(\frac{CS}{2\,x\,(4.2\,x\,10^{-9})}\right)^2 + \frac{[SO_2]}{(4.2\,x\,10^{-9})}(8.6\,\mathrm{x}\,10^{-9}\,x\,GlobRad + 6.1\,\mathrm{x}\,10^{-29}[O_3][Alkene])\right]^{1/2}$$

The results from predicting sulphuric acid from the testing data sets using the above equation are shown in Figure R1 below, and the results from predicting sulphuric acid from 10 000 different *k* value combinations specific to the site are shown in figure R2. Note that the 10 000 different *k* value combinations refer to the 10 000 iterations performed on each time step including bootstrap resampling and accounting for predictor biases. Complementary error analyses to figure R2A are shown in figure R2B. The detailed method used to determine the *k* value combinations from the training data set, as well as the one obtained from the equation above, are explained in details in the previous section. We also show the model prediction error which was estimated as the interquartile range of the modelled H_2SO_4 concentration of a single point in time arising from the uncertainty in *k* values for each of the sites.

Moreover, we verified the four fits on the testing data set; i.e. the full Equation 2, the equation without the Stabilized Criegee Intermediates source (Equation 4), the equation without the cluster sink term (Equation 5) and the equation without neither the Stabilized Criegee Intermediates source nor the cluster sink term (Equation 6). We found that Fit 1 (Full equation) best defines the measured sulphuric acid concentration in comparison to the rest with a high correlation coefficient between the measured and the modelled data (R = 0.70) and a slope of 0.997 (Figure R3). The diurnal cycle is also nicely described by the Equation 4 which captures both nighttime and daytime (Figure R4).

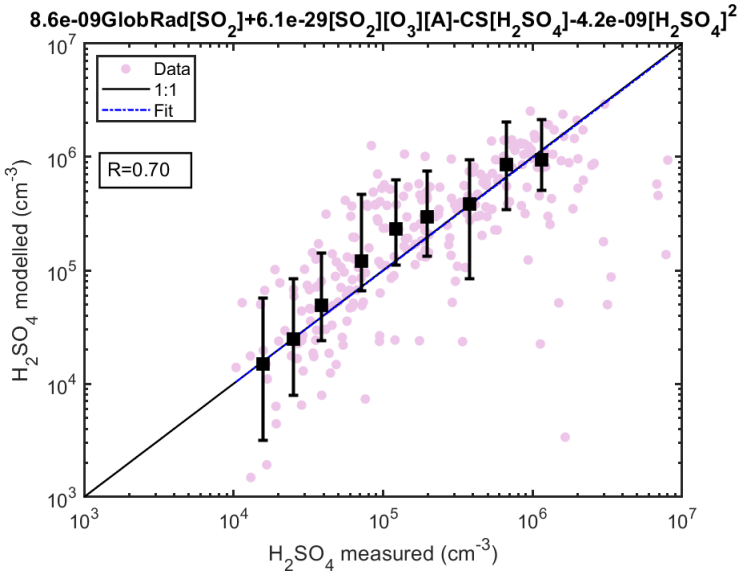


Figure R 1 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Hyytiälä SMEAR II station. The concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. The modelled concentrations are derived using equation 9. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations.

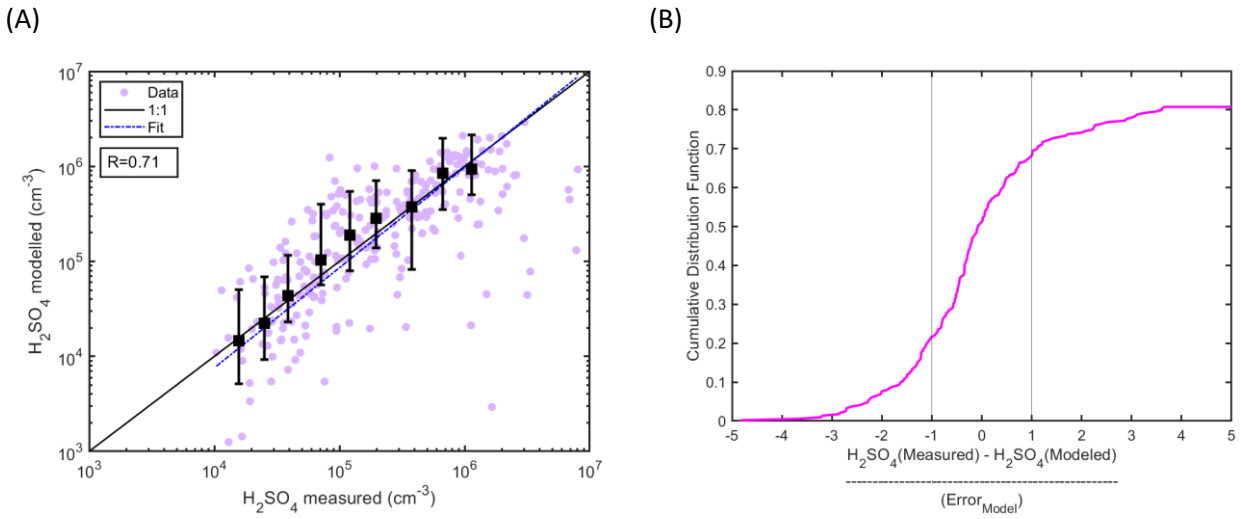


Figure R 2 (A) Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Hyytiälä SMEAR II station. The concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations. (B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 257 data points).

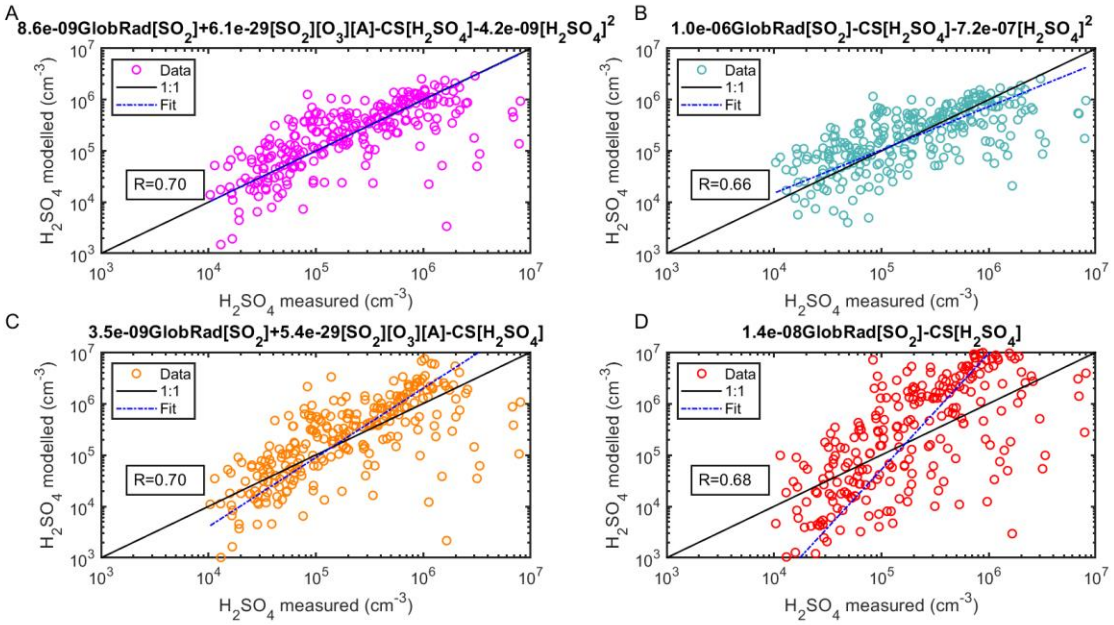


Figure R 3 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at SMEAR II station, Hyytiälä Finland using the four different combinations of source and sink terms. The concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration.

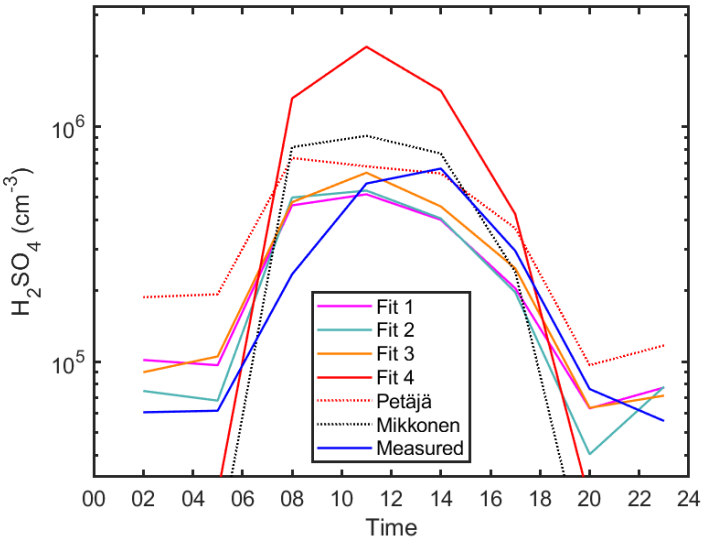


Figure R 4 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR II in Hyytiälä, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in (Petäjä et al., 2009)(Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

1.2. Rural location: Agia Marina Equation 10 (Glob Rad >= 50).

$$[H_2SO_4]_{rural} = -\frac{CS}{2\ x\ (2.2\ x\ 10^{-9})} + \left[\left(\frac{CS}{2\ x\ (2.2\ x\ 10^{-9})}\right)^2 + \frac{[SO_2]}{(2.2\ x\ 10^{-9})}(9.7\ \mathrm{x}\ 10^{-8}\ x\ GlobRad)\right]^{\frac{1}{2}}$$

An additional location 'Helsinki', representative of a semi-urban location was introduced for testing the predictive power of the rural proxy equation. Note that the rural equation was chosen over the urban equation, since the CS and SO_2 concentrations measured in Helsinki matched those in Agia Marina (rural location) rather than those in Budapest (urban location); see Figure 10 (ACPD). For testing the predictive power of the rural background site proxy (Equation 10), we used measurements from July 1, 2019 to July 16, 2019 during daytime (GlobRad >= 50 W/m^2). Results show that although the modelled sulphuric acid concentrations did not correlate as well as in other locations (R = 0.44), the bias could be attributed to the missing source (alkene) in the original equation as mentioned in the previous section. Indeed, looking at the binned data, we found that at within each concentration bin, the modelled sulphuric concentrations tend to span the 1:1 line. Actually, the discrepancy between the measured and the modelled concentration was smaller than the model prediction error (Figure R6). Note that the model prediction error was estimated as the interquartile range of the modelled H2SO4 concentration of a single point in time arising from the uncertainty in k values. For the rural background site, we also found that the diurnal cycle is better described when introducing the additional clustering sink term (Figure R7).

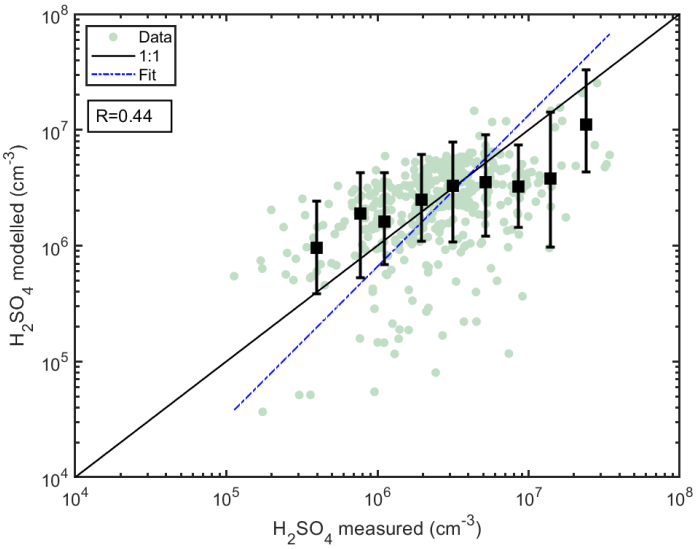


Figure R 5 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Helsinki SMEAR III station. The concentrations shown are 1-hour medians resulting in a total of 416 data points. The modelled concentrations are derived using equation 10. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations.

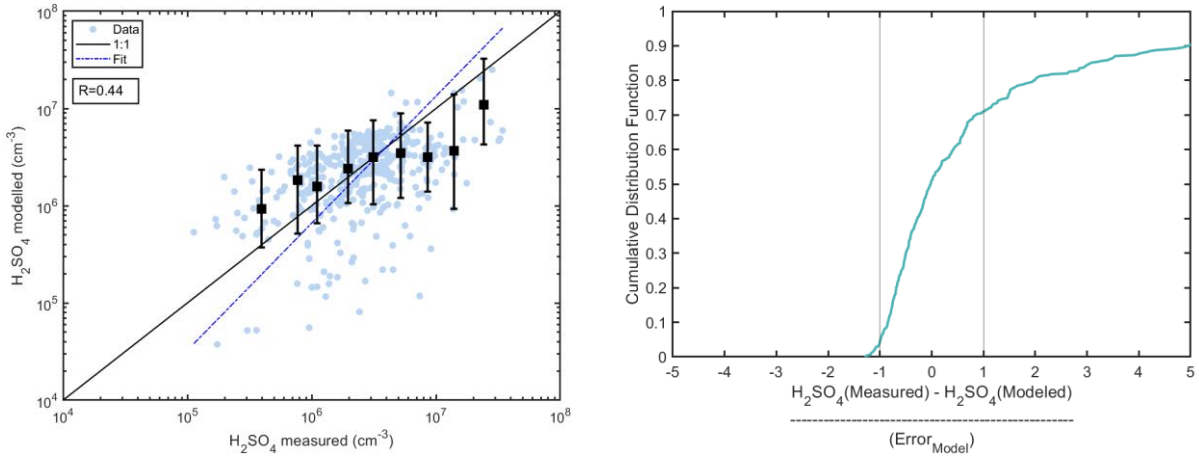


Figure R 6 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Helsinki SMEAR III station. The concentrations shown are 1-hour medians resulting in a total of 416 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations.(B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 416 data points).

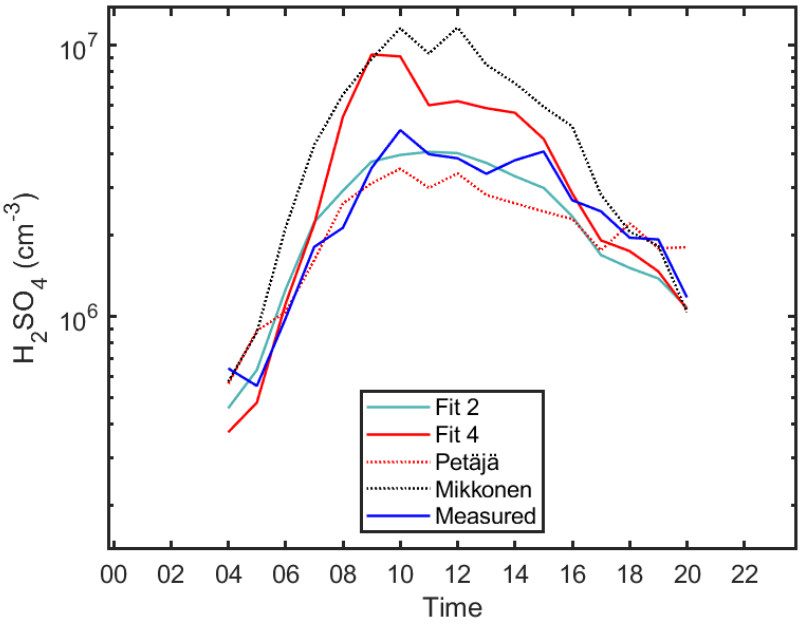


Figure R 7 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR III in Helsinki, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

1.3. Megacity: Beijing: Equation 12.

$$[H_2SO_4]_{megacity} = -\frac{CS}{2\ x\ (7.0\ x\ 10^{-9})} + \left[\left(\frac{CS}{2\ x\ (7.0\ x\ 10^{-9})}\right)^2 + \frac{[SO_2]}{(7.0\ x\ 10^{-9})}(1.94 \times 10^{-8}\ x\ GlobRad + 1.44 \times 10^{-29}[O_3][Alkene])\right]^{1/2}$$

We applied the equation on an additional independent data set from the same location between September 8, 2019 and October 15, 2019. The results show that the modelled sulphuric acid concentrations correlated well (R = 0.84) with the measured sulphuric concentrations, with a slope of ~1.1 for the testing data set (Figure R8). Also for this site, we tested the four fits on the testing data set; i.e. the full Equation 2, the equation without the Stabilized Criegee Intermediates source (Equation 4), the equation without the cluster sink term (Equation 5) and the equation without neither the Stabilized Criegee Intermediates source nor the cluster sink term (Equation 6). We found that Fit 1 (Equation 4) best defines the measured sulphuric acid concentration in comparison to the rest of the equations. The results show a high correlation coefficient between the measured and the modelled data (R = 0.84) and a slope of 1.03 (Figure R10). The diurnal cycle is also nicely described by the Equation 4 which captures both nighttime and daytime (Figure R11).Similar to the boreal forest and rural site predictions, in Beijing, the discrepancy between the measured and the modelled concentration is also smaller than the model prediction error (Figure R9).

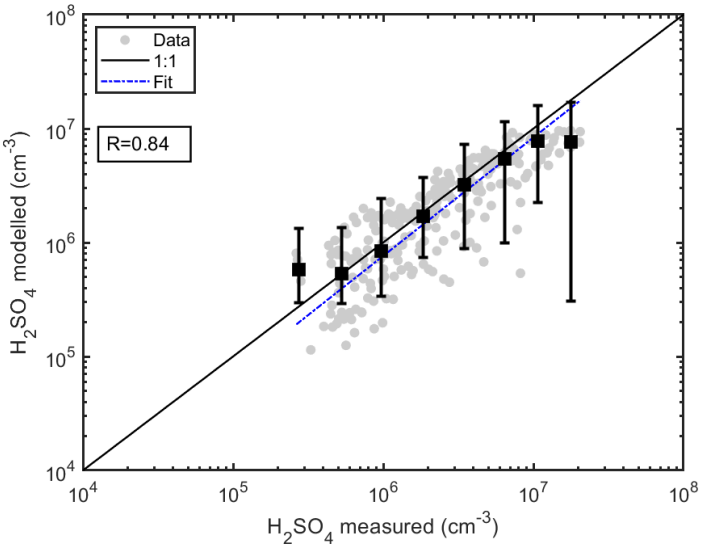


Figure R 8 Sulphuric acid concentrations modelled as a function of measured sulphuric acid in Beijing. The concentrations shown are 1-hour medians resulting in a total of 268 data points. The modelled concentrations are derived using equation 12. The gray data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations.

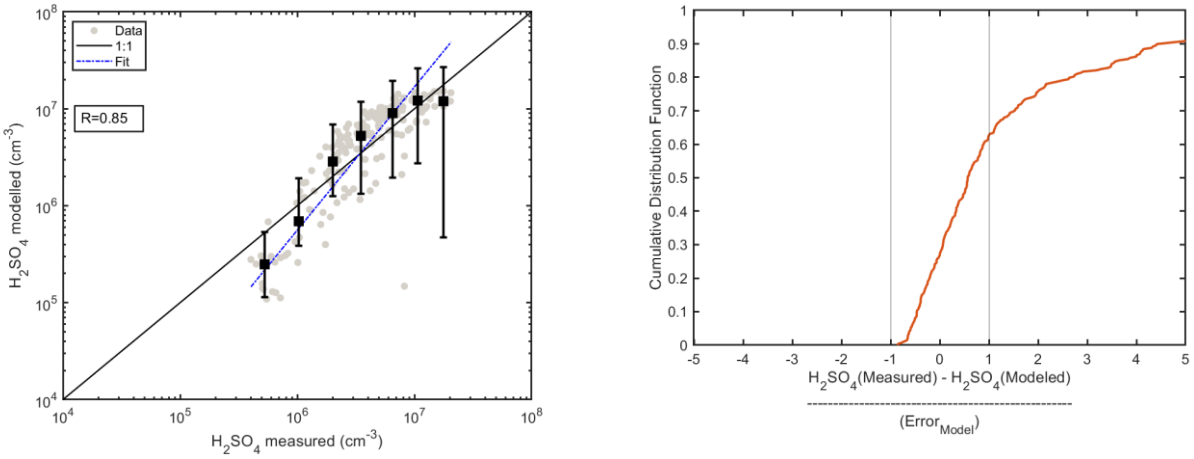


Figure R 9 Sulphuric acid concentrations modelled as a function of measured sulphuric acid in Beijing. The concentrations shown are 1-hour medians resulting in a total of 263 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The gray data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whiskers correspond to 25th and 75th percentiles of the predicted concentrations.(B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 263 data points).

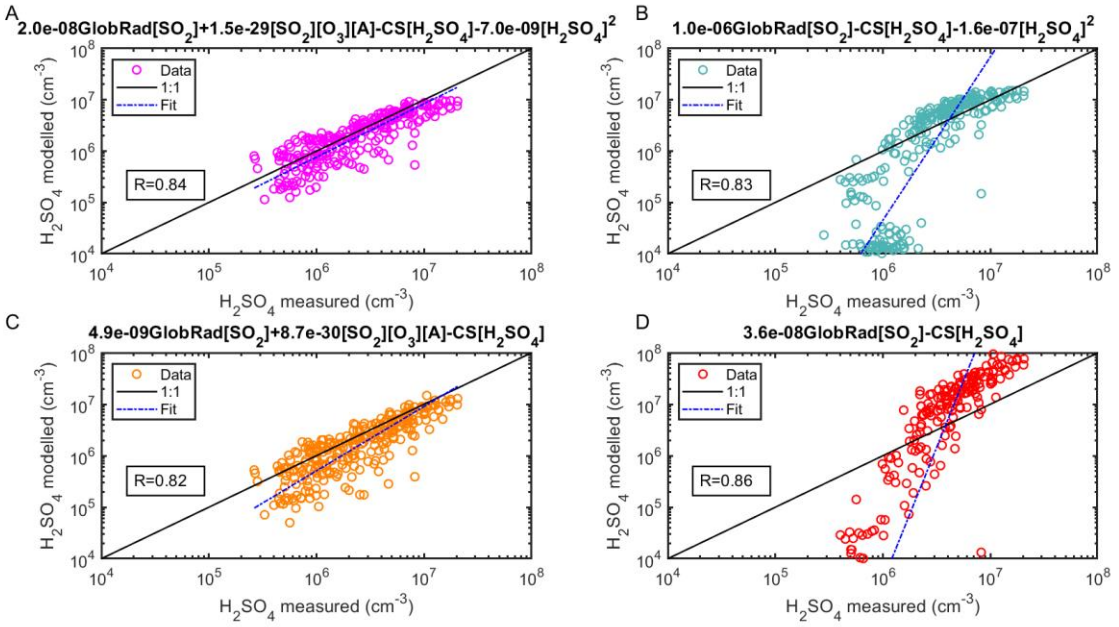


Figure R 10 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at SMEAR II station, Hyytiälä Finland using the four different combinations of source and sink terms. The concentrations shown are 1-hour medians resulting in a total of 263 data points in each subplot. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

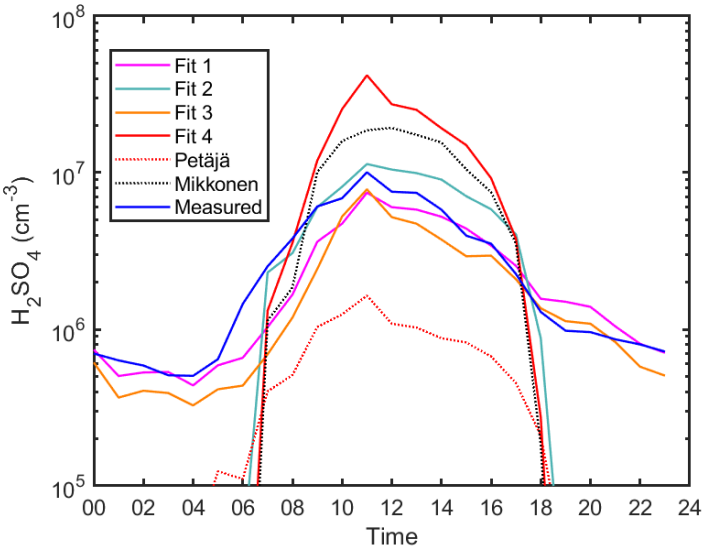


Figure R 11 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at in Beijing, China for the testing data set. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

1.4 Kilpilahti: Equation 10

Finally, we did a very interesting test where we tested the predictive power of our developed proxy on a data set measured at an industrial area in close proximity to an oil refinery. Interestingly, the median CS at the location lies within the interquartile range of the CS measured in Hyytiälä and that measured in Agia Marina. The SO_2 concentrations at the measurement site were higher than in both Hyytiälä and Agia Marina, but smaller than the ones reported in Budapest. Additionally, we observed that alkene concentrations at Kilpilahti were within the range of those monitored in Hyytiälä, which is attributed to the green belt in the area (Sarnela et al., 2015). Accordingly, we tested the proxy equation 9 on the Kilpilahti data set. Our results showed that Equation 9 derived for Hyytiälä is able to predict the sulphuric acid concentrations in Kilpilahti with a high correlation coefficient (R= 0.74) (Figure R12). Similar to other locations, the Fit 1 (Equation 4) best describes the sources and sinks at the location (Figure R14). The discrepancy between the measured and the modelled concentration is smaller than the model prediction error for less than 50% of the data points only (Figure S13). This observation is consistent with the diurnal cycle (Figure R15). During certain mornings (4:00 – 8:00 LT), when the measured sulphuric concentrations were particularly high, the model was unable to predict the concentrations accurately. These high concentrations were attributed to air masses coming from the oil refinery (Sarnela et al., 2015). Indeed, our proxy was not able to explain these morning peaks using biogenic alkenes, however, in such an industrial area, anthropogenic sources could play a role in determining the magnitude of sulphuric acid concentrations. With the condensation sink being rather low (median ~0.005 s^{-1}), the impact of direct H_2SO_4 emissions cannot be ruled out either.

.

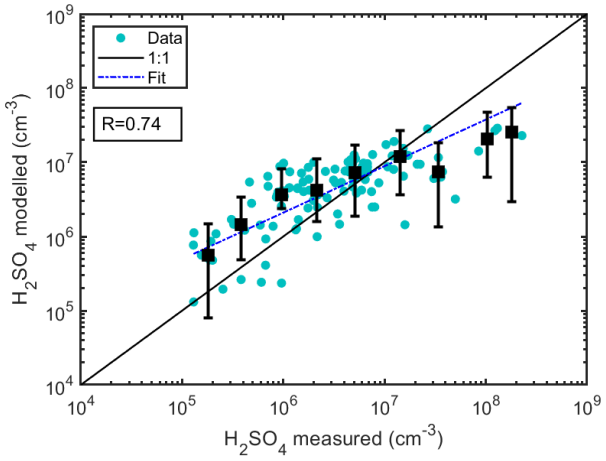


Figure R 12 Sulphuric acid concentrations modelled as a function of measured sulphuric acid. The colored data points refer to the modelled (predicted) concentrations at Kilpilahti Finland, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whiskers correspond to 25th and 75th percentiles of the predicted concentrations. The concentrations shown are 1-hour medians resulting in 114 data points. The modelled concentrations are derived using equation 9.

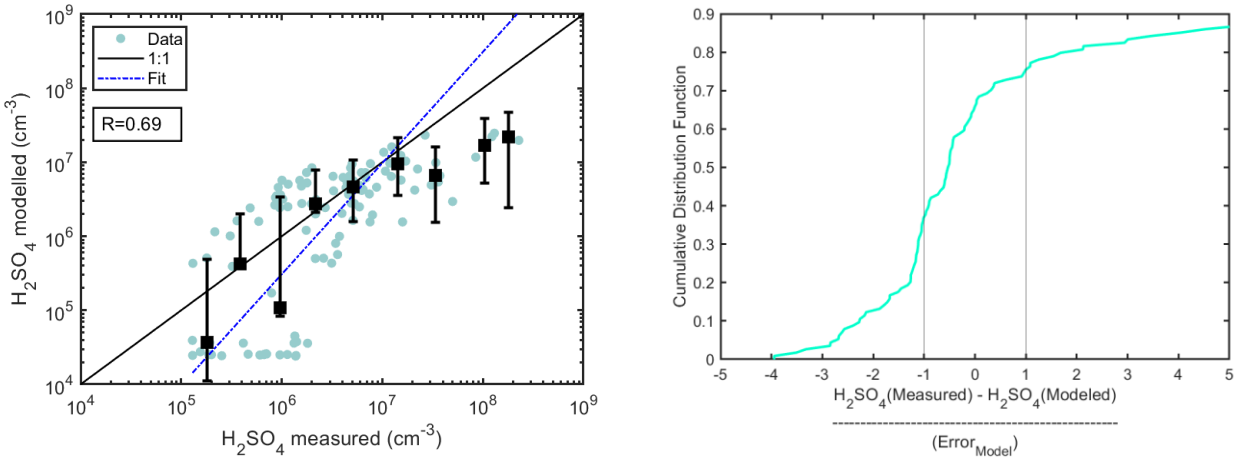


Figure R 13 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Kilpilahti, Finland. The concentrations shown are 1-hour medians resulting in a total of 114 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the the boreal forest location. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations. (B) Cumulative distribution function of the model error weighted difference between measured and modeled H2SO4 concentration (using 114 data points).

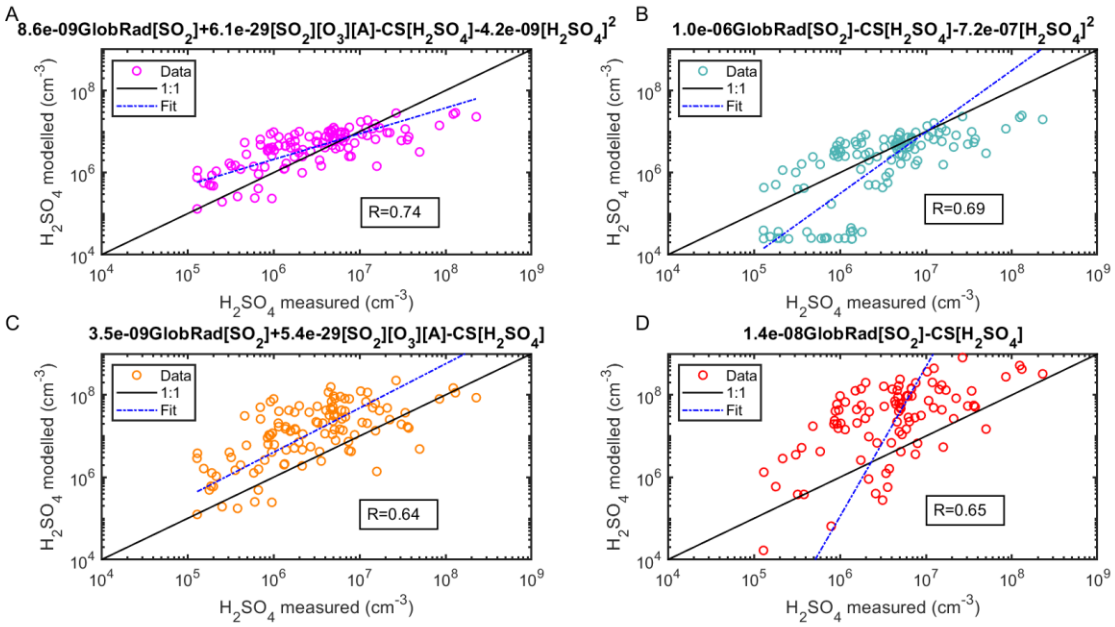


Figure R 14 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at Kilpilahti, oil refinary Finland using the four different combinations of source and sink terms derived from Hyytiälä. The concentrations shown are 1-hour medians resulting in a total of 114 data points in each subplot. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

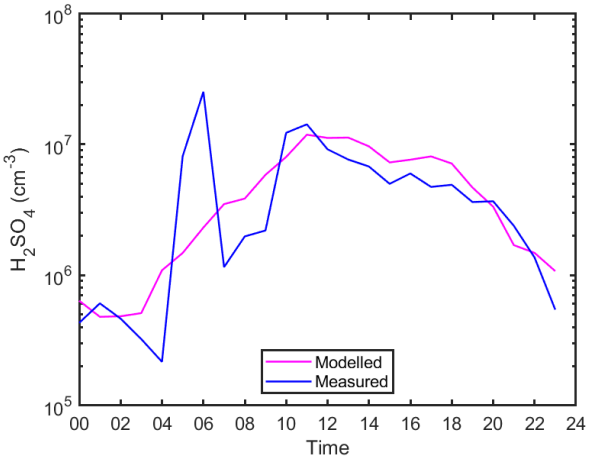


Figure R 15 The diurnal variation of sulphuric acid proxy concentrations observed concentrations at Kilpilahti, industrial area, Finland. Median values are shown. The modelled concentration is predicted using Equation 9 using the k values derived from Hyytiälä SMEAR II station.

2. Monthly variation of the sources and sinks in both Hyytiälä and Beijing

Since our paper tackles mostly the sources and sinks of H_2SO_4 in various locations and not only aims at deriving a physical proxy and in order to assess the representative qualities of the data sets we used, we included monthly variation of the sources and sinks in both Hyytiälä and Beijing during which we have extended data sets which include nighttime calculations (Figure R16).
The text on Line 401 now reads:

The Criegee intermediate term showed its importance mostly when global radiation is low, not only in nighttime but also during winter (Figure 11) in both Hyytiälä and Beijing.

And on Line 414:

The cluster term is found to contribute most during spring daytime in Hyytiälä (Figure 12 – A & C), which is the time window during which clustering and thus new particle formation events happen (Dada et al., 2018; Dada et al., 2017) The same is observed for Beijing, where the clustering term contributed up to 70% of the total sink terms during daytime (Figure 12-D).

Additionally, we added a paragraph describing the representative nature of our data sets in comparison to the whole year for all site by comparing to available literature from each site.

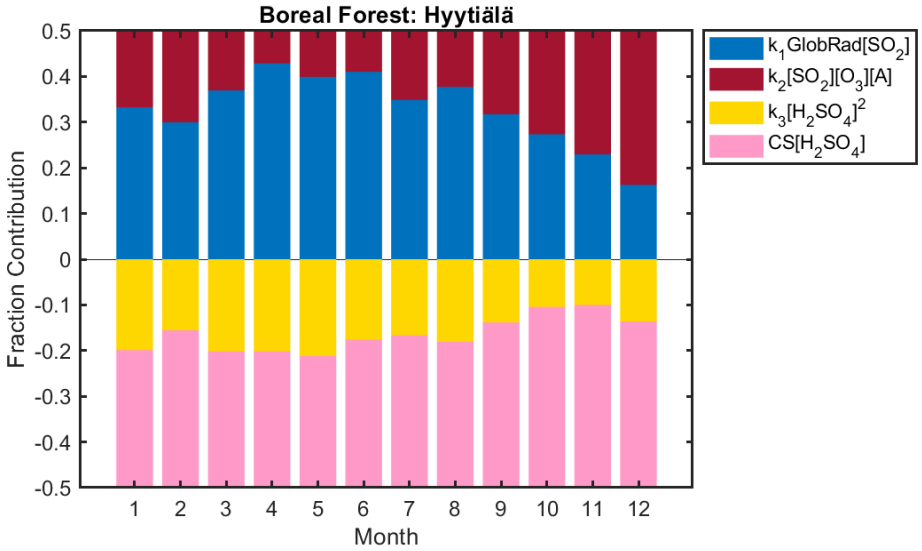
The text on Line 141 now reads:

Trace gases measured during the short campaign periods in Agia Marina and Budapest are representative of yearly concentrations in respective locations when compared to longer term measurements at the same site (Salma et al., 2016; Baalbaki, 2020, In Prep.).

and on Line 155:

Condensation sink values obtained during the short campaign periods in Agia Marina, Helsinki and Budapest are representative of yearly concentrations in respective locations when compared to longer term measurements at the same site (Salma et al., 2016; Baalbaki, 2020, In Prep.).

(A)



(B)

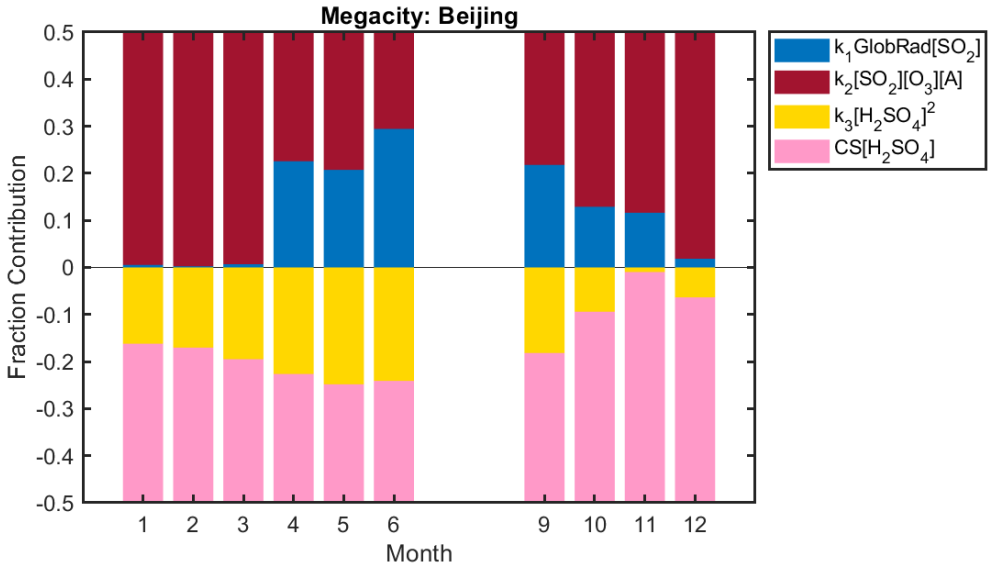


Figure R 16 (A) Monthly variation of each source and sink term to the change in H_2SO_4 concentration in Hyytiälä during the period of the training data set 2016-2019 (excluding 2017). (B) Monthly variation of each source and sink term to the change in H_2SO_4 concentration in Beijing using a combined data set between January and December 2019. The data outside the training and testing data set has missing measured sulphuric acid concentration and proxy concentrations were used in obtaining this figure.

3. Derivation of night-time proxies in Hyytiälä and Beijing

We agree with the reviewer that the sources of the sulphuric acid may shift between day and night hours. Indeed, during dark hours, the Criegee intermediates' source is dominant. However, we think that extent of the contribution of each source term depends on the concentration of the precursor vapour rather than on the *k* itself, where k could be temperature dependent resulting in a difference between day and night. Nevertheless, we did the analysis for day and night separately. We compared the results from the separate (day and night) analysis to those from considering one equation as in Figure R 17.

First, we found that a better fit between the measured and training data set proxy concentrations is found when using one equation for daytime and nighttime than for daytime alone which has to

do with the different points in time. Additionally, we found that the *k* values derived from 10 000 iterations for all day, daytime and nighttime separately have distinct characteristics (Figure R19). First, k_1 values derived from all day, daytime alone or nighttime alone are within the range of each other. Interestingly, the k_2 values for daytime or nighttime alone are also similar, while when fitting one equation for daytime and nighttime together the k_2 values show different character. This means that separating the equation into day and night independently would depict the pattern of the predictor in this case the alkene term (Figure R20). The alkene term has a strong diurnal and seasonal cycle as shown in figure R20.

We performed the same analysis on the Beijing data set after we reassessed the Global Radiation data. In order to perform the 4 fits on any data set, the global radiation cannot be zero as otherwise Fit 2 fails completely. Therefore, in the case of Beijing we set the global radiation zero values into half the minimum observed radiation, which is assumed to be equivalent to the detection limit of the instrument ($GlobRad_{min}$ = 0.03 W/m^2). After reassessing the global radiation data, we came to the same conclusion as for Hyytiälä, which is that one single equation for daytime and nighttime together is capable of explaining the sulphuric acid concentrations without Beijing biased to the diurnal or seasonal pattern of any of the predictor variables. The only obstacle was that when fitting one bulk equation for daytime and nighttime together unconstrained, the fit resulted in an unphysical k_3 value of the order of 0.01. In order to overcome this, we restricted the upper limit of the k_3 value to the median we get from fitting daytime data only. This assumption is acceptable since clustering is dominant during daytime. Indeed, when we then compared the daytime alone fits versus the ones from the bulk equation, we observed a better fit (Figure R21-R22). Additionally, different k_1 values for daytime and nighttime were obtained when fit separately, in general during the nighttime the global radiation is too low, and therefore has too low variability and therefore for this parameter the nighttime is poorly defined, which explains why the k_1 in this condition is an order of magnitude higher. When we fitted the data together, the k_1 matches the one from the daytime, which is not poorly defined. Therefore, also for Beijing we fitted the daytime and nighttime together (Figure R23). All in all, we think that introducing the predictive power of each of the equations, as suggested by the reviewer, was an excellent idea which helped in assessing whether using a bulk equation is enough for either location. Indeed, as shown in the previous section, for Hyytiälä the bulk proxy equation serves well in predicting both nighttime and daytime concentrations of sulphuric acids during the independent data set period. Similarly, obtaining the bulk equation from the spring time Beijing training data was able to predict both nighttime and daytime concentrations during summer and autumn in Beijing during the testing data set period.

However, in order to show the difference between daytime and nighttime in terms of sources or sinks, we decided to show diurnal contribution of those for both Hyytiälä in Beijing (Figure R 24-25). Similar to the observations from the monthly cycles, the diurnals show that when the global radiation is available the sulphuric acid formation pathway rather goes through the SO_2- OH mechanism. During dark hours, the Criegee pathway dominates the sulphuric acid source. Additionally, clustering is dominant during daytime hours. Please see insertions to the main text in the section above.

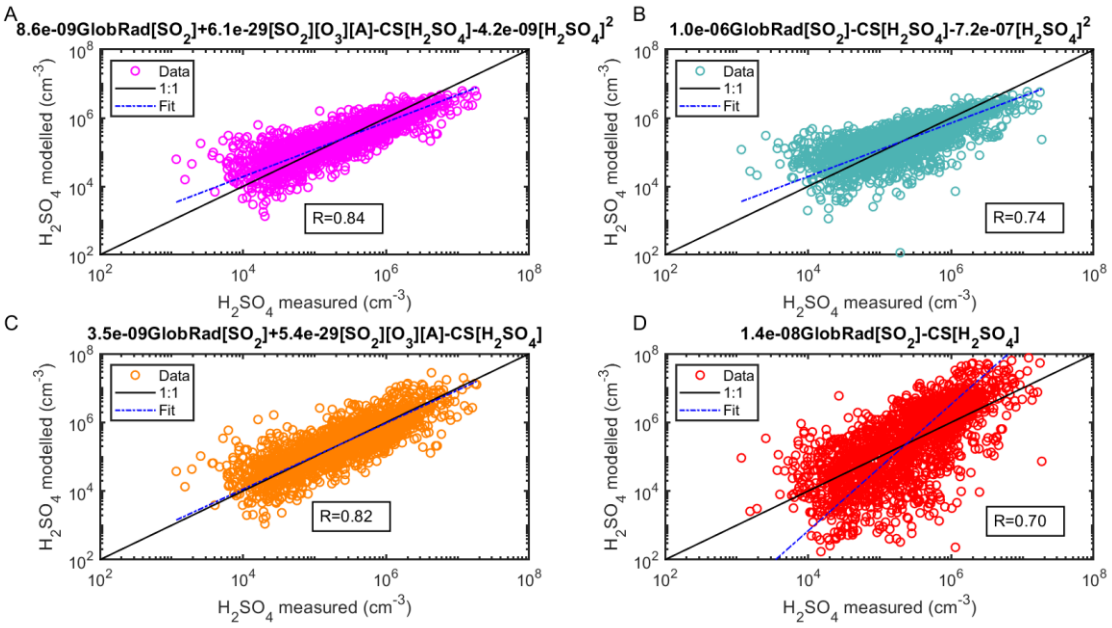


Figure R 17 Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at SMEAR II station, Hyytiälä Finland. The observed concentrations are measured 2016-2019 using CI-APi-ToF and are 3-hour medians resulting in a total of 1860 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration.

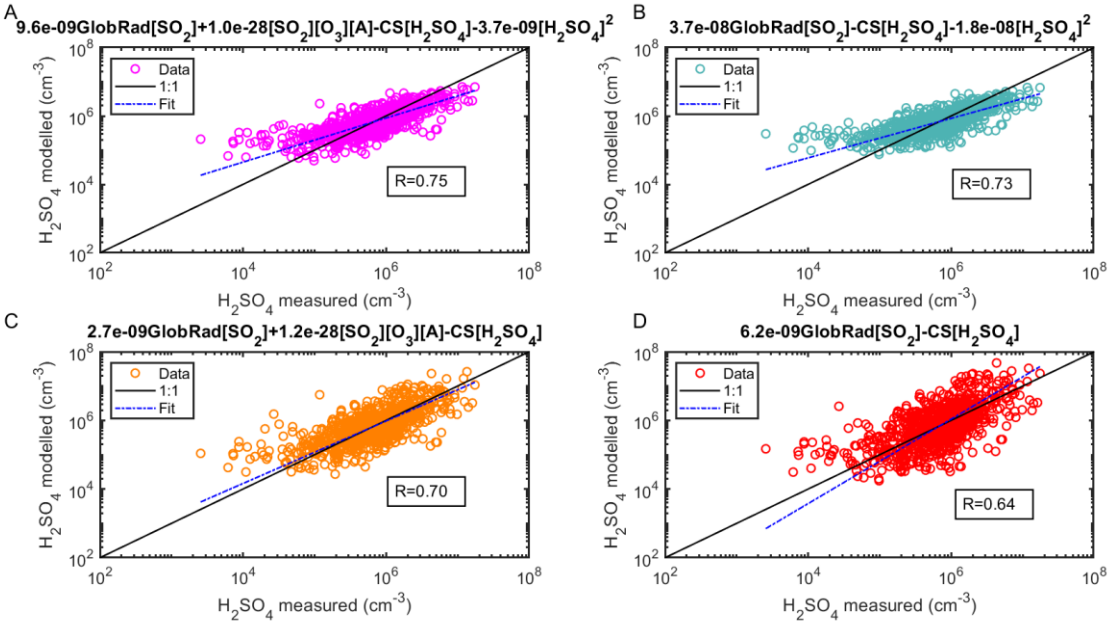


Figure R 18 Sulphuric acid proxy concentration as a function of measured sulphuric acid ***during daytime*** *(GlobRad >= 50 W/m²). Observation at SMEAR II station, Hyytiälä Finland. The observed concentrations are measured 2016-2019 using CI-APi-ToF and are 3-hour medians for daytime data resulting in a total of 921 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration.*

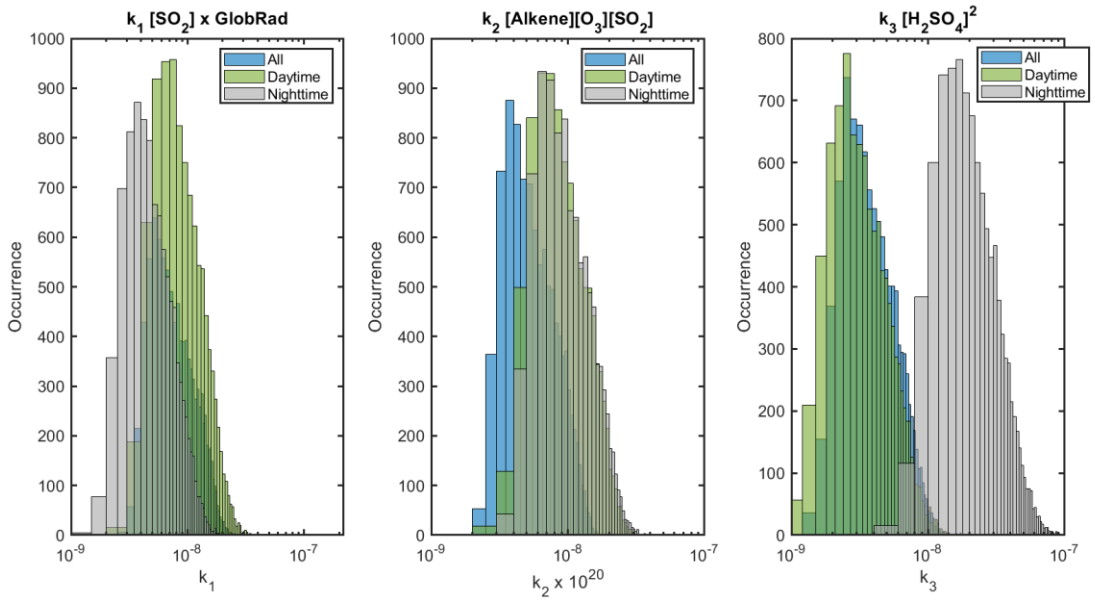


Figure R 19 Histograms showing the occurence of k values derived from 10,000 disturbed booststrap resampling runs when fitting a full-day proxy denoted by 'All' and colored in blue, a daytime proxy denoted by 'Daytime' and colored in green, and a nighttime proxy denoted by 'Nighttime' and colored in grey.

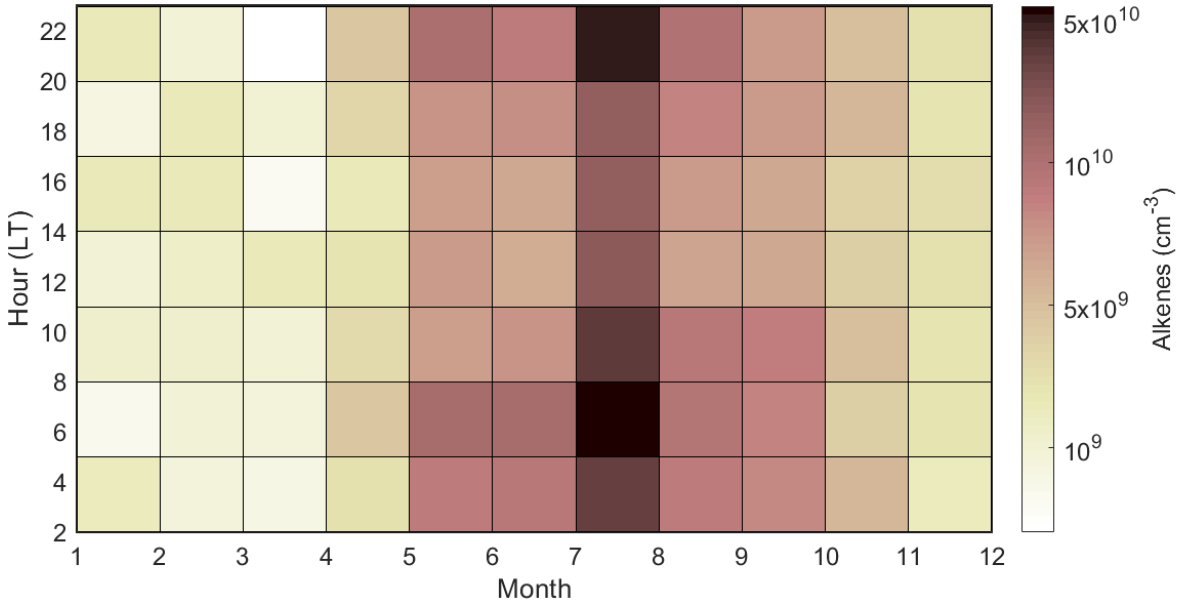


Figure R 20 Temporal variation in the median monoterpene concentration in Hyytiälä 2016- 2019. Observation at SMEAR II station, Hyytiälä Finland. The observed concentrations are measured 2016-2019 using PTR-ToF, see also Perakyla et al. (2014).

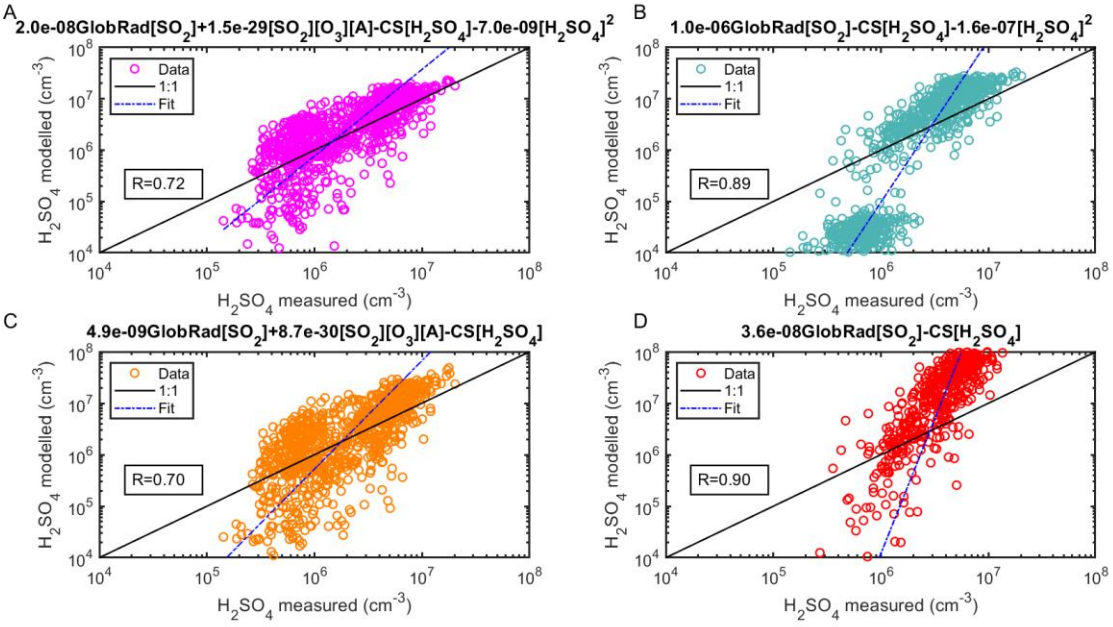


Figure R 21 (A) Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at Beijing, China. The observed concentrations of the training data set are measured in 2019 using CI-APi-ToF and are 1-hour medians resulting in a total of 877 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). Coefficients shown on top of the subplots relate to the daytime values. The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b). Note that the upper limit of the cluster term k value is limited to the same value as the daytime value to avoid getting unphysical values which were observed (k_3 = 0.01) in case no limit on the k value is added.

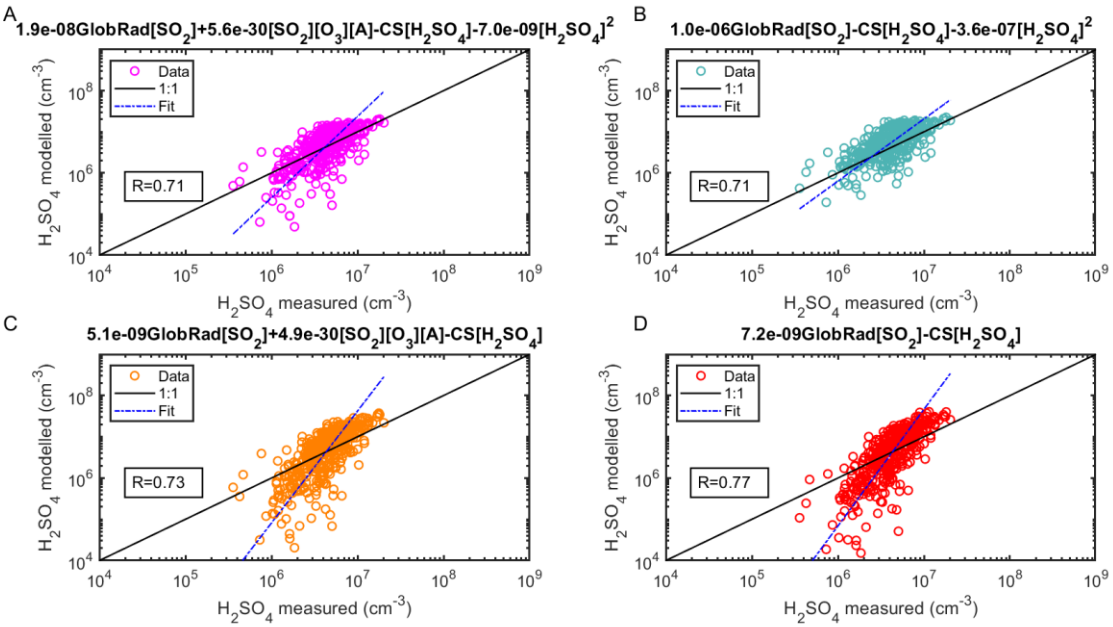


Figure R 22 (A) Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at Beijing, China during daytime GlobRad >= 50 W/m². The observed concentrations of the training data set are measured in 2019 using CI-APi-ToF and are 1-hour medians resulting in a total of 415 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). Coefficients shown on top of the subplots relate to the daytime values. The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

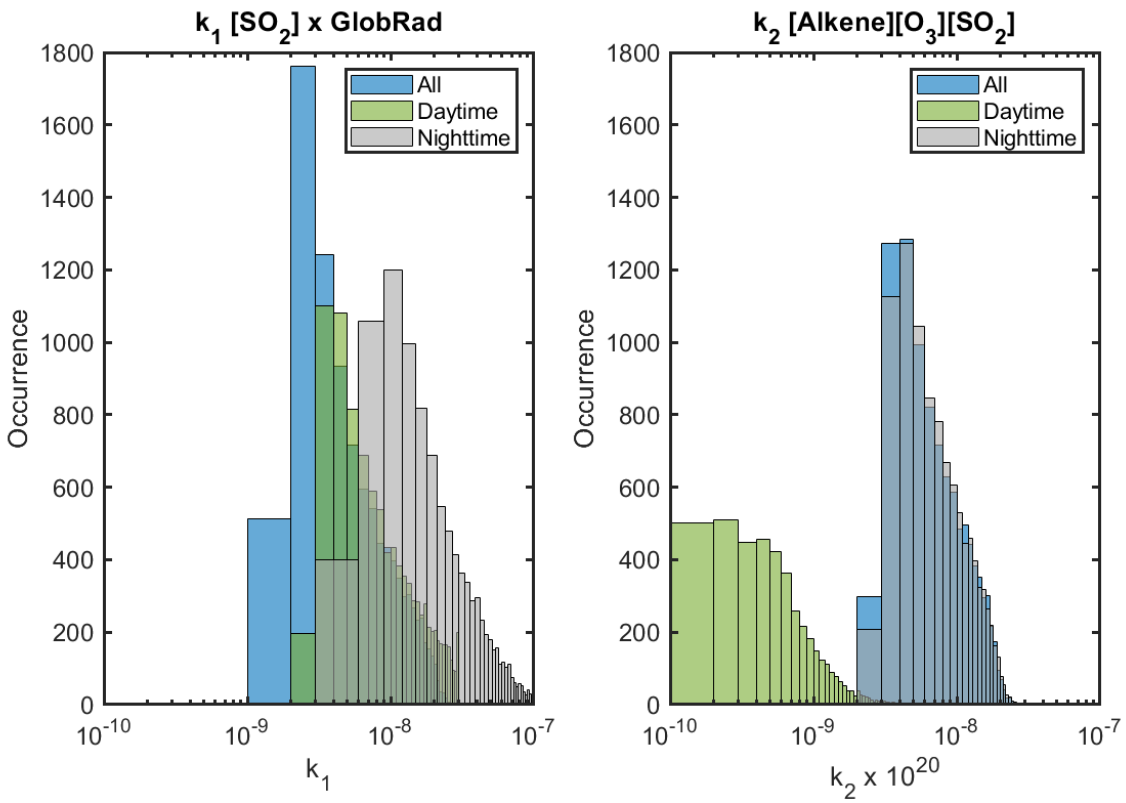


Figure R 23 Histograms showing the occurrence of k values derived from 10,000 disturbed bootstrap resampling runs when fitting a full-day proxy denoted by 'All' and colored in blue, a daytime proxy denoted by 'Daytime' and colored in green, and a nighttime proxy denoted by 'Nighttime' and colored in grey in Beijing. Note that the k_3 values are not shown since they are similar to the daytime values due to limiting the k_3 to the upper limit of the daytime k_3 value.

.

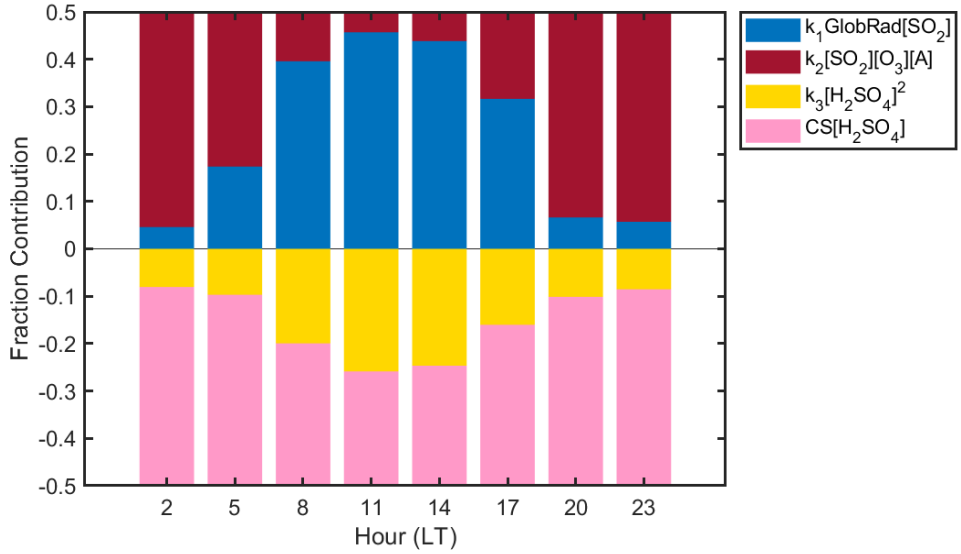


Figure R 24 Diurnal variation of each source and sink term to the change in H_2SO_4 concentration in Hyytiälä within the training data set.

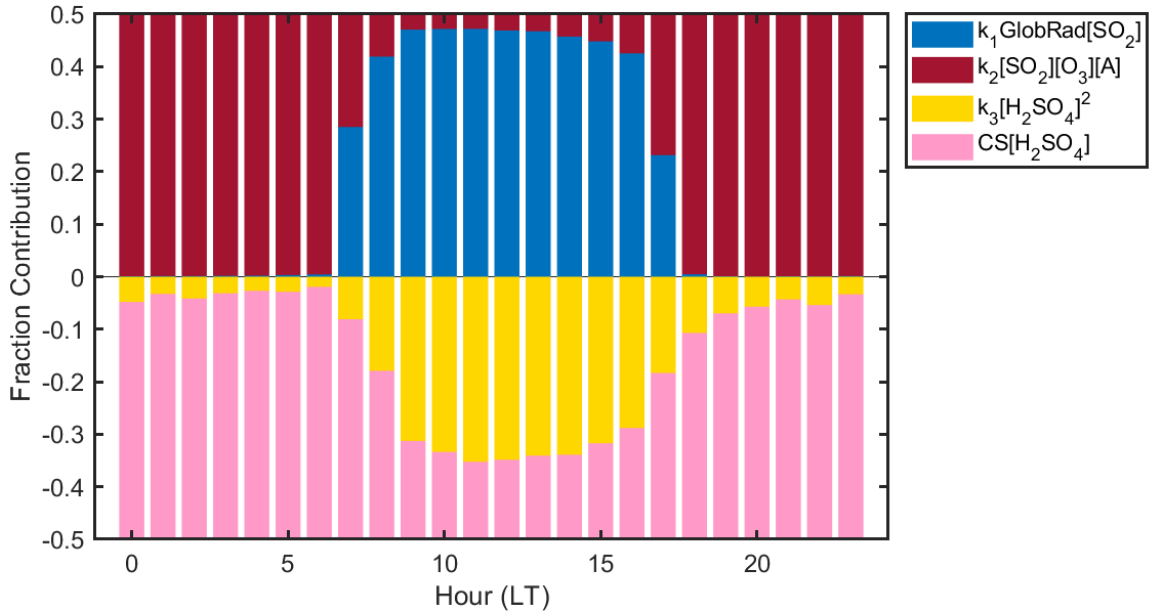


Figure R 25 Diurnal variation of each source and sink term to the change in H_2SO_4 concentration in Beijing within the training data set.

Specific comments:

Point-by-point replies to the specific comments are added below.

1. Page 2 line 76: proved->suggested

Modified.

2. Page 2 lines 91-93: Bold statements, considering the comments in this revision regarding generalizability

Modified.

In order to evaluate our hypothesized sources and sinks and derive the proxy equations, we utilize measurements from four different locations: (1) Hyytiälä, Finland, (2) Agia Marina, Cyprus, (3) Budapest, Hungary and (4) Beijing, China, representing a semi-pristine boreal forest environment, rural environment in the Mediterranean area, urban environment and heavily polluted megacity, respectively. To evaluate the predictive power of the derived proxies, the equations are further tested on independent data sets. We further compare the coefficients of production and losses in each environment in order to understand the prevailing mechanism of the H_2SO_4 budget in each of the studied environments. As a result of this investigation, a well-defined sulphuric acid concentration can be derived for multiple areas around the world and even extended in time during times when it was not measured (such as: gap filling, forecast, prediction, estimation, etc.).

3. Page 3, lines 102-104: Were all the measurements made on the same platform?

Measurements of different variables within the same location are performed at the same platform except for Hyytiälä and Helsinki. We added details related to the measurement platforms of every variable to section 2.2.

4. Page 3, lines 130-134: I have recently learnt that calibrating CI-APi-ToF is not an easy task (Talk by Ylisirniö et al. EAC2019). Were the instruments calibrated such that the results between sites are comparable and are the measured concentrations of realistic magnitude?

We agree that different organic compounds calibrations are still mystery (Talk by Ylisirniö et al. EAC2019), however, calibrations of sulphuric acid are straightforward and robust. The instruments in all four locations were calibrated in a similar way using the method presented by (Kurten et al., 2012) and the results are comparable.

We added the following to the Line 134:

In all locations, the CI-APi-ToF instruments were calibrated in a similar way prior to the campaign using the method presented by Kurten et al. (2012) to ensure the results are comparable.

5. Page 4, lines 145-155: CS was reported in Hyytiälä with RH correction and in other sites no such correction is defined. The CS measures should be consistently defined if the results are being generalized.

We agree with the reviewer that including a hygroscopic growth correction for only the boreal forest results in a discrepancy when inter-comparing. Therefore, we reassessed the fits for the boreal forest location using condensation sink values calculated in the same way as in the rest of the studied locations. The results of this fit would be suitable for comparing the sources and sinks in various

locations. We replaced the equation and related k values in the main text with those reassessed, see figure R26.

However, we think that the boreal forest environment has been studied thoroughly over the years and it is ideal to use the best data we have and all the information we could. In case any of the readers is interested in calculating a sulphuric acid proxy from Hyytiälä, we recommend that they use the equation which includes corrected CS for hygroscropic growth.

In fact, we found that the fit with the hygroscopic correction is better than that without this correction. See figure R17 (no correction) in comparison to figure R26 (with correction).

The results and equations are added into the supplementary information and the related text in main text. Line 302 now reads:

Furthermore, we derived an additional proxy equation using CS corrected for hygroscopic growth (Laakso et al., 2004) to be used when calculating a more robust proxy for Hyytiälä. The details, equation and results are shown in the supplementary information (Figure S10-S12).

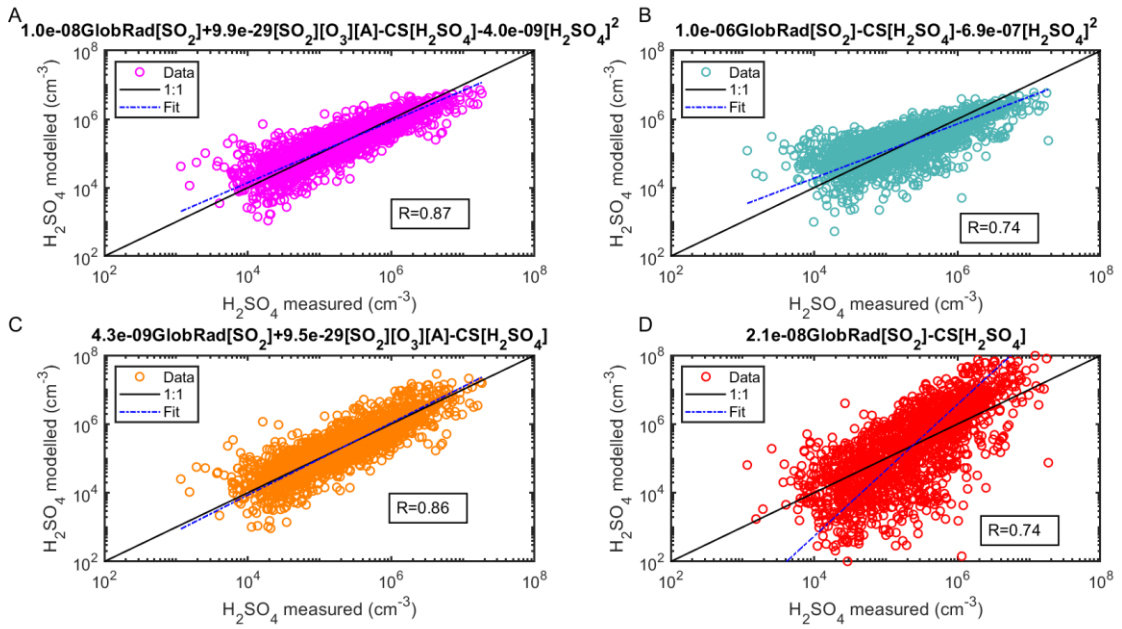


Figure R 26 Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at SMEAR II station, Hyytiälä Finland ***with CS corrected for hygroscopic growth****. The observed concentrations are measured 2016-2019 using CI-APi-ToF and are 3-hour medians resulting in a total of 1594 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration(log(y)=a.log(x)+b).*

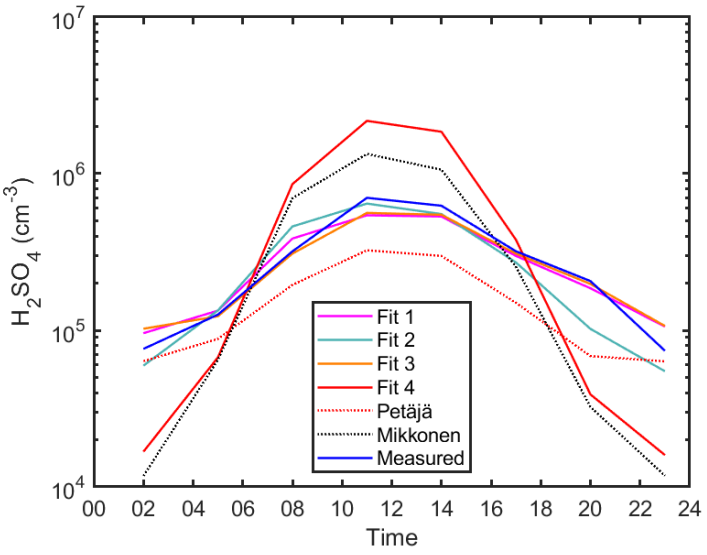


Figure R 27 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR II in Hyytiälä, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

6. Page 5, lines 183-184 and Figures S3-S7: Why Pearson correlation coefficients? The data are most probably not normally distributed and they contain outliers, which violate the basic assumptions of Pearson correlation.

The reviewer is right. We used the scatter plots between the variables to decide which coefficient we should use. We replaced the Pearson with a Spearman coefficients in Figures S2-S6.

7. Page 5, lines 203-209: How the sink term $k_3[H2SO4]^2$ is defined? It needs to be clarified here for usability of the proxy.
8. Pages 5-6, Equations: Overall, the notation of the equations is somewhat confusing. First term is clear, does the second term refer similarly as the first one that it is k_2 times ozone concentration times Alkene concentration times SO2 concentration? In addition, does [H2SO4] in third term refer to sulphuric acid concertation or that the CS is calculated for sulphuric acid? Does in last term [H2SO4] 2 refer to squared concentration, and if yes, drawn from where? I suggest clarification of the equations.

As per the suggestion of the two previous comments a clarification has been added to the text to explain the 3rd and 4th terms of the Equation 1

$$\frac{d[H_2SO_4]}{dt} = k_0[OH][SO_2] + k_2[O_3][Alkene][SO_2] - CS[H_2SO_4] - k_3[H_2SO_4]^2 \qquad (1)$$

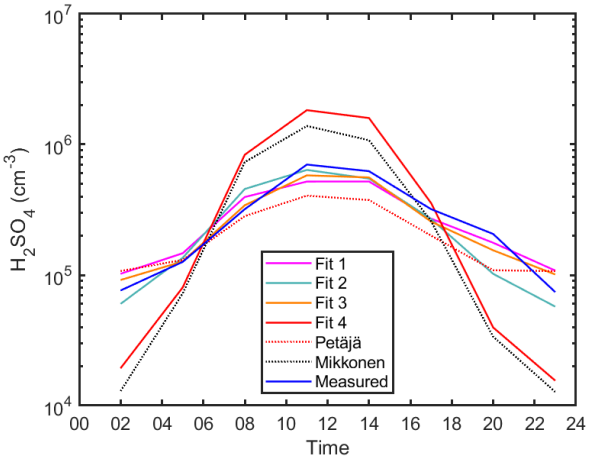
The text on line 201 now read:

The third term in Equation 1 represents the loss of H_2SO_4 into pre-existing aerosol particles, known as condensation sink (CS) and is calculated by multiplying the CS calculated for sulphuric acid with the concentration of sulphuric acid monomer. The fourth term in Equation 1 is defined as the square of sulphuric acid concentration multiplied by clustering coefficient k_3. The square of sulphuric acid represents the collision of two sulphuric acid monomers forming a sulphuric acid dimer, which was found to be the first step of atmospheric cluster formation (Yao et al., 2018). Therefore, this term takes into account the additional loss of H_2SO_4 due to cluster formation not included in the term containing CS. This is necessary because CS is only inferred from size-distribution measurements at maximum down to 1.5 nm, i.e. not containing any cluster concentrations and hence losses onto these clusters. This term is written in the form of sulphuric acid dimer production, which seems to be the first step of cluster formation once stabilized by bases (Kulmala et al., 2013; Almeida et al., 2013; Yao et al., 2018).

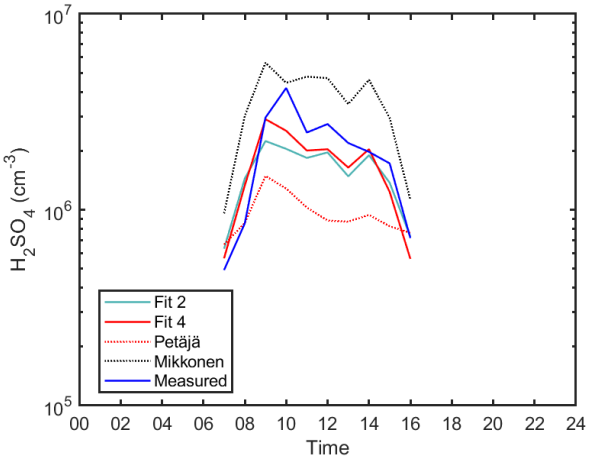
9. Page 6, lines 242-249: It is not surprising to see that the Petäjä proxy had some difficulties, as it is constructed only with data from Hyytiälä. Already in Mikkonen et al. (2011) it was seen that the Petäjä proxy is not always working well outside of Hyytiälä. Thus, it would be interesting to see comparisons on proxy from Mikkonen et al., which has been shown to work in varying environments.

We compared our proxies with Mikkonen et al. 2011 in all 4 locations, and added the diurnal Mikkonen plot to the main text (Figures 2,4, 6 and 8) while the scatter plots between measured sulphuric acid concentrations and both of Petäjä and Mikkonen proxies during daytime (GlobRad >= 50 W/m^2) in Figures S13 and S14, respectively.

Hyytiälä



Agia Marina



Budapest

Beijing

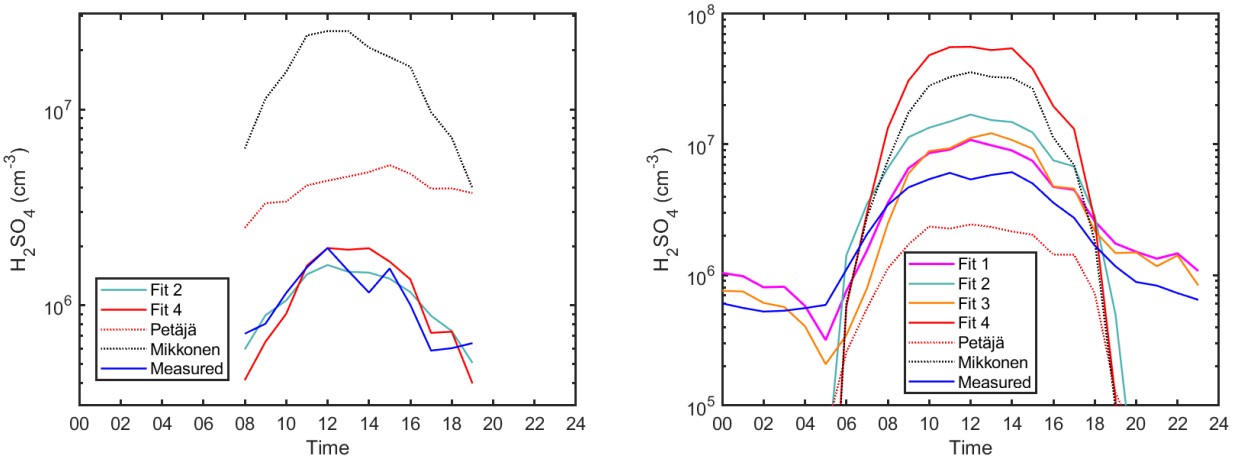


Figure R 28 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7) and Mikkonen et al. 2011 (Equation 8).

Hyytiälä

Agia Marina

Budapest

Beijing

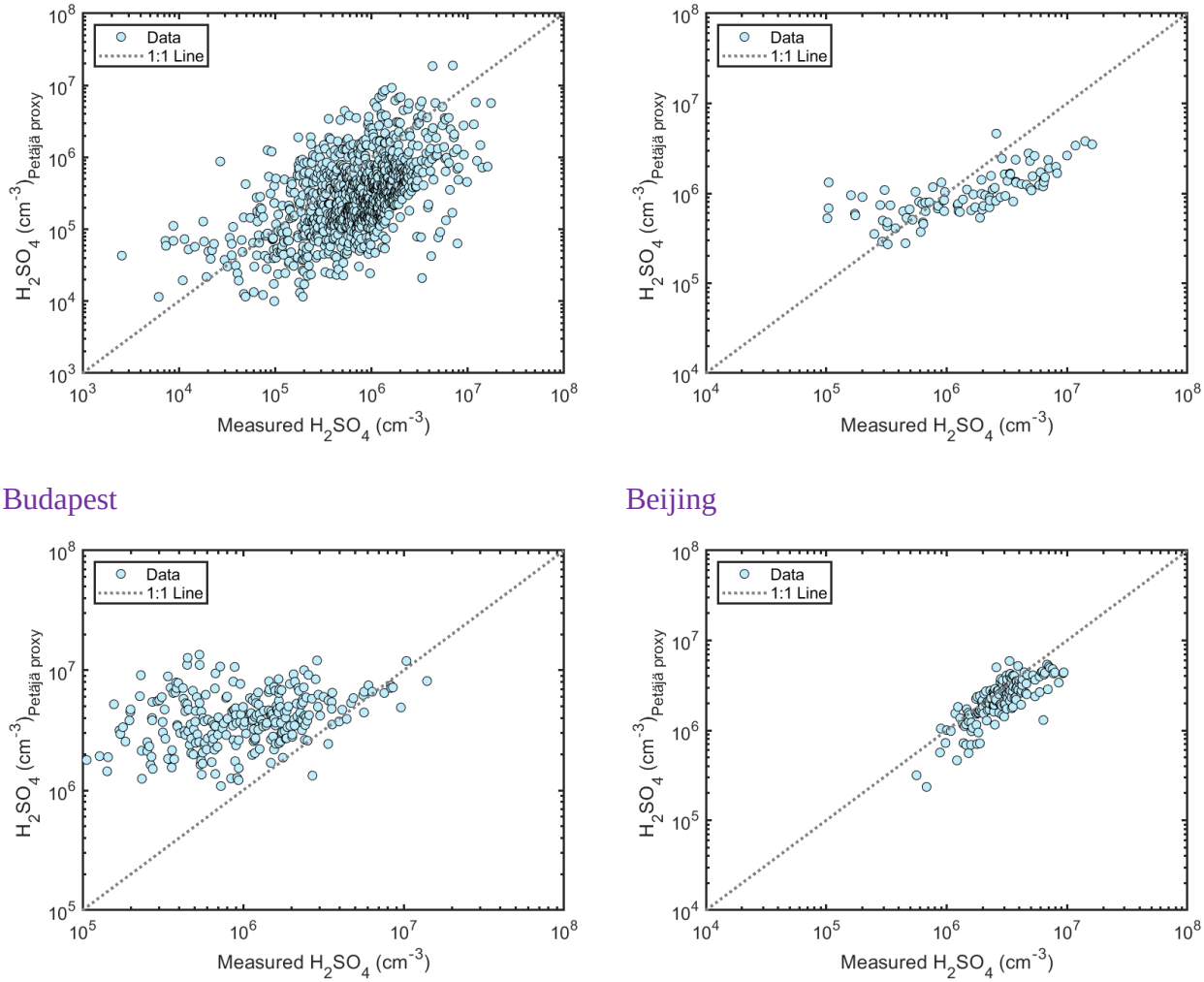
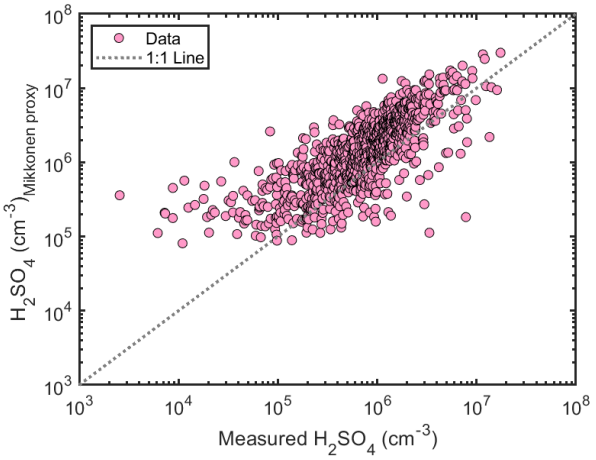
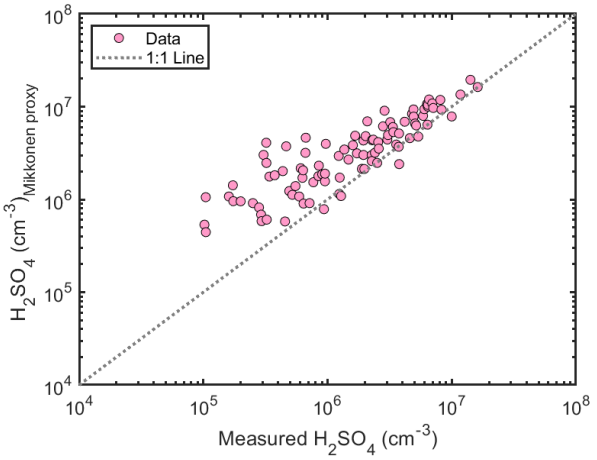


Figure R 29 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Petäjä et al. 2009 proxy at the 4 locations during daytime (GlobRad >= 50 W/m2): Hyytiälä, Agia Marina, Budapest and Beijing.

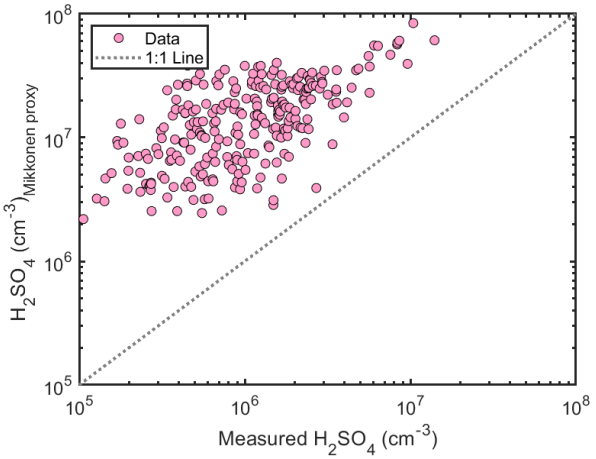
Hyytiälä



Agia Marina



Budapest



Beijing

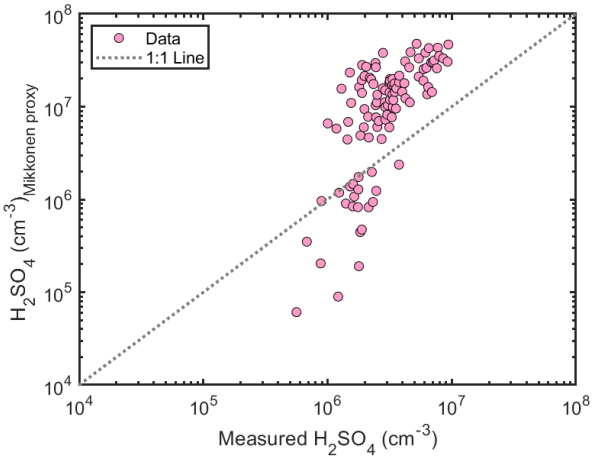


Figure R 30 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Mikkonen et al. 2011 proxy at the 4 locations during daytime (GlobRad >= 50 W/m²): Hyytiälä, Agia Marina, Budapest and Beijing.

10. Page 6 lines 251-254: The predictor variables in the proxy contain high measurement uncertainty. Does the fminsearch procedure take that account?
11. Page 6 lines 254-257: I am happy to see uncertainty estimation for the coefficients made with bootstrap! Though some details on bootstrap procedure should be provided, e.g. how many resamples were drawn?

Answers to question 10 and 11 are added to the beginning of this document.

12. Page 6 lines 260-265: How does the AIC reflect the probability of over- or under-fitting in these analyses? As calculating log-likelihood for AIC might be sensitive for number of observations was it checked that the N was the same for all proxies in certain site? With multiple instruments in use, there might be gaps in data indifferent time points.

The reviewer is right that the AIC criterion is sensitive or even driven by the N. In order to avoid the bias due to number of observation points per fit, we selected the data points when all variables are

available simultaneously. We also add a Table S4 which shows the parameters included in deriving the AIC in each site. See also next comment.

13. Page 7, line 273 and Figure 1: Are the numbers of data points the same in each subplot?

For each location separately, all the subplots contain the same number of points. Although it might be possible to include more points in the panels where no alkene term is included, yet for comparability reasons, especially for the AIC we kept a constant number of data points per subplot. The number of points to each of the subplot for all 4 locations is shown in the corresponding figure caption. A table S4 describing the statistics included in the AIC calculation such as the number of points, correlation coefficients, slope .. etc. is added to the supplementary information.

14. Figure 2 and related text in chapter 4.1: Do I read the figure correctly that the proxy values from 23-02 are missing? If this is due to missing global radiation, this could be corrected by the suggestion above to derive separate night-time proxy.

There is no missing data except that the PTR measurements for alkenes are every 3 hours. We are sorry for the typo in the figure 1 caption. Now it is corrected.

15. Page 7, line 308: "…proves the truthfulness…" is quite an overstatement

We agree with the reviewer that using the same data set for deriving and predicating is not a valid method for a proxy derivation. Besides adding a complete section on the predictive powers of the derived proxies, we modified the above sentence into:

The correlation between the measured and proxy concentration of H_2SO_4 was 0.88 (96 data points) which shows that the chosen predictors were able to explain the measured sulphuric acid concentration largely (Figure 3).

16. Figure 5: Why the scale is from 10^2 when the data starts from 10^5? Overall, the observed concentrations seem rather low for urban environment. Were the conditions somewhat unusual during the measurement campaign?

The figure is fixed. Concerning the overall concentrations, we do not think that there were any unusual conditions. The measured concentrations are within the range of observations between Hyytiälä and Beijing. We added a time series of the measured H_2SO_4 in Budapest in the supplementary information (Figure S1) to help show the variation in the H_2SO_4 concentrations upon changes in meteorology.

17. Page 9, lines 388-389: Clarify how the predicted fractions were drawn for table 2 and fig 9

Line 389 now reads:

The contribution of the various source and sink terms to the change of H_2SO_4 concentrations are determined using Equation 2. The median derived k_1, k_2 and k_3 values, together with the measured H_2SO_4, CS, trace gases and GlobRad per site, were used to calculate each of the terms. Source term 1 refers to k_1 x GlobRad x $[SO_2]$, source term 2 refers to k_2 x $[O_3]$ x [Alkene] x $[SO_2]$, sink term 3 refers to k_3 x $[H_2SO_4]^2$ and sink term 4 refers to CS x $[H_2SO_4]$. The contribution of each term is then calculated as the median or percentiles of the normalized term to the sum of all terms.

18. Table 2: 27th percentile?

This was a typo, we changed it to 75th.

19. Figure 10: Global radiation distribution is missing. The basic statistics could also be given in (supplement) table. Sulphuric acid concentration in Megacity seems also low.

Global radiation distribution and a table of basic statistics was added to the supplementary information.

20. Page 10, lines 438-440: It is stated that the coefficients did not vary substantially, I might disagree. But regardless of that, did you try to pool the data from different sites an calculate a combined data proxy? Naturally with Equation 4 which could be calculated for all sites. Would this give a more generalizable proxy?

We agree with the reviewer that unifying the parametrization with the aim of coming up with 1 equation would be nice. In this sense, we unified the day and night time equations wherever possible and present now unified equations each for Beijing and Hyytiälä. These equations perform well in explaining the diurnal variability at the respective site. Unifying wasn't possible for the Cyprus and Budapest datasets because of missing alkene data. Merging Hyytiälä and Beijing to come up with a single proxies would require accounting for different alkene mixes (boreal forest dominated by biogenic VOCs, Beijing strongly impacted by anthropogenic VOCs). And yes, we revisited our k values, illustrated in Figure R31, the k_2 related to the sulphuric acid formation through Criegee intermediates is clearly different at both locations. Additionally, with the different sizes of data sets from each of the locations, when we tried to assess one parametrization using Equation 4, as suggested with the reviewer, the fit was bias to the Hyytiälä data which has the highest contribution. Therefore, we opt here not to further unify, yet agree with the reviewer that such efforts should be targeted in future results together with distinguishing further chemical processes such as the contribution of different VOC classes.

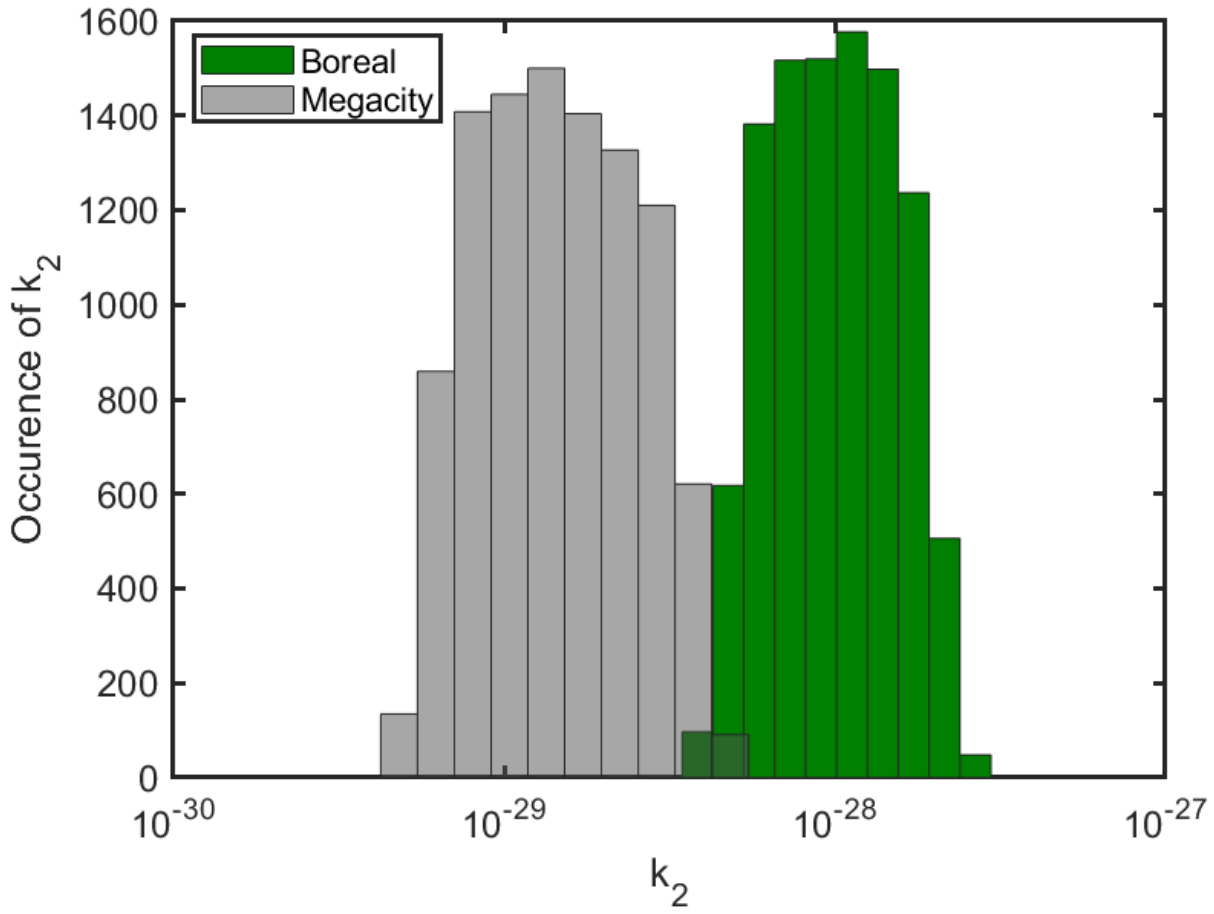


Figure R 31 Histogram showing the distribution of k2 values from 10,000 iterations in both Hyytiälä and Beijing.

21. Discussion and suggestions section: It would be helpful to give here the direct equations for calculating the proxies in each site. It would probably increase the future use of the derived proxies. Equations could also be an appendix.

The equations 9-12 are added to Table 1.

References

Almeida, J., Schobesberger, S., Kurten, A., Ortega, I. K., Kupiainen-Maatta, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurten, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppa, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petaja, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipila, M., Stozhkov, Y., Stratmann, F., Tome, A., Trostl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P. L., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamaki, H., and Kirkby, J.: Molecular understanding of sulphuric acid-amine particle nucleation in the atmosphere, Nature, 502, 359-363, 10.1038/nature12663, 2013.

Baalbaki, R., Pikridas M., Jokinen T., Dada L., Ahonen L., Lehtipalo K., Petäjä T., Sciare J. Kulmala M.: Towards understanding the mechanisms of new particle formation in

the Eastern Mediterranean, 2020, In Prep.

Dada, L., Paasonen, P., Nieminen, T., Mazon, S. B., Kontkanen, J., Perakyla, O., Lehtipalo, K., Hussein, T., Petaja, T., Kerminen, V. M., Back, J., and Kulmala, M.: Long-term analysis of clear-sky new particle formation events and nonevents in Hyytiala, Atmos Chem Phys, 17, 6227-6241, 10.5194/acp-17-6227-2017, 2017.

Dada, L., Chellapermal, R., Buenrostro Mazon, S., Paasonen, P., Lampilahti, J., Manninen, H. E., Junninen, H., Petäjä, T., Kerminen, V. M., and Kulmala, M.: Refined classification and characterization of atmospheric new-particle formation events using air ions, Atmos. Chem. Phys., 18, 17883-17893, 10.5194/acp-18-17883-2018, 2018.

Efron, B., and Tibshirani, R. J.: An introduction to the bootstrap, CRC press, 1994.

Kulmala, M., Kontkanen, J., Junninen, H., Lehtipalo, K., Manninen, H. E., Nieminen, T., Petaja, T., Sipila, M., Schobesberger, S., Rantala, P., Franchin, A., Jokinen, T., Jarvinen, E., Aijala, M., Kangasluoma, J., Hakala, J., Aalto, P. P., Paasonen, P., Mikkila, J., Vanhanen, J., Aalto, J., Hakola, H., Makkonen, U., Ruuskanen, T., Mauldin, R. L., Duplissy, J., Vehkamaki, H., Back, J., Kortelainen, A., Riipinen, I., Kurten, 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, 10.1126/science.1227385, 2013.

Kurten, A., Rondo, L., Ehrhart, S., and Curtius, J.: Calibration of a chemical ionization mass spectrometer for the measurement of gaseous sulfuric acid, Phys Chem A, 116, 6375-6386, 10.1021/jp212123n, 2012.

Laakso, L., Petaja, T., Lehtinen, K. E. J., Kulmala, M., Paatero, J., Horrak, U., Tammet, H., and Joutsensaari, J.: Ion production rate in a boreal forest based on ion, particle and radiation measurements, Atmos Chem Phys, 4, 1933-1943, DOI 10.5194/acp-4-1933-2004, 2004.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E.: Convergence Properties of the Nelder--Mead Simplex Method in Low Dimensions, SIAM Journal on Optimization, 9, 112-147, 10.1137/s1052623496303470, 1998.

Perakyla, O., Vogt, M., Tikkanen, O. P., Laurila, T., Kajos, M. K., Rantala, P. A., Patokoski, J., Aalto, J., Yli-Juuti, T., Ehn, M., Sipila, M., Paasonen, P., Rissanen, M., Nieminen, T., Taipale, R., Keronen, P., Lappalainen, H. K., Ruuskanen, T. M., Rinne, J., Kerminen, V. M., Kulmala, M., Back, J., and Petaja, T.: Monoterpenes' oxidation capacity and rate over a boreal forest: temporal variation and connection to growth of newly formed particles, Boreal Environ Res, 19, 293-310, 2014.

Petäjä, T., Mauldin Iii, R., 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, 10.5194/acp-9-7435-2009, 2009.

Salma, I., Németh, Z., Weidinger, T., Kovács, B., and Kristóf, G.: Measurement, growth types and shrinkage of newly formed aerosol particles at an urban research platform, Atmos. Chem. Phys., 16, 7837-7851, 10.5194/acp-16-7837-2016, 2016.

Sarnela, N., Jokinen, T., Nieminen, T., Lehtipalo, K., Junninen, H., Kangasluoma, J., Hakala, J., Taipale, R., Schobesberger, S., Sipila, M., Larnimaa, K., Westerholm, H., Heijari, J., Kerminen, V. M., Petaja, T., and Kulmala, M.: Sulphuric acid and aerosol particle production in the vicinity of an oil refinery, Atmos Environ, 119, 156-166, 10.1016/j.atmosenv.2015.08.033, 2015.

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, 10.1126/science.aao4839 2018.

Review of Dada et al "Sources and sinks driving sulphuric acid concentrations in contrasting environments: implications on proxy calculations" by Anonymous Referee

The manuscript, "Sources and sinks driving sulphuric acid concentrations in contrasting environments: implications on proxy calculations," by Dada et al. describes a new method for estimating gas phase H2SO4 concentrations using relatively common measurements. The development of these so-called "proxies" for H2SO4 is important as this species is often used in global models for simulating the timing and intensity of new particle formation events. Additional proxies are especially needed for representing regions that were not include in previous attempts (e.g., China) or during time periods that we not considered previously (e.g., nighttime). Thus, this manuscript is potentially valuable and is, in principle, worthy of publication in ACP. I do however, wish to point out a one main item and a few minor issues that I would like the authors to respond to prior to recommending publication.

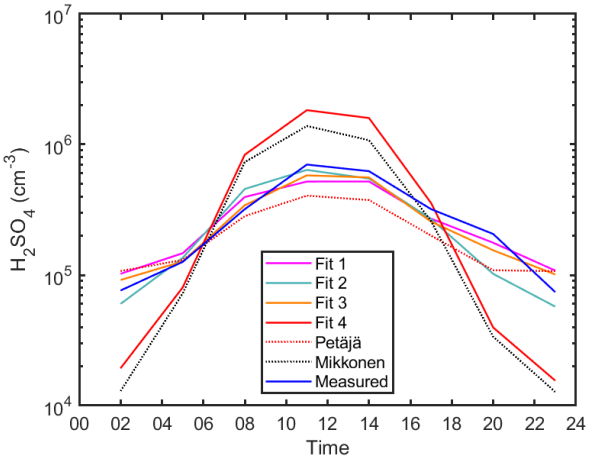
We thank the reviewer for their valuable comments and suggestions, we think that these help improve the presentation of the proxy and the overall quality of the study. We provided point-by-point answers in purple. Insertions to the text are in *Italics*. Line numbers refer to the old version of the ACPD version of the text.

As a major concern: In the abstract of this manuscript and throughout the text the authors claim that the new proxy is "a more flexible and an important improvement of previous proxies." While that may be true, we only are provided a comparison to the previous proxy developed in a pristine boreal forest atmosphere (the Petaja proxy). Nowhere do the authors compare their new proxy to that developed by Mikkonen et al. First of all, this makes little sense as the Mikkonen model was developed for a broader range of conditions than the Petaja model. If there is a valid reason to disregard the Mikkonen model then the authors should state that, or else they should show model predictions from that on all relevant figures as they did with the Petaja model. Otherwise they should remove the statement that the model is an improvement over other proxies, as they are only comparing to one.

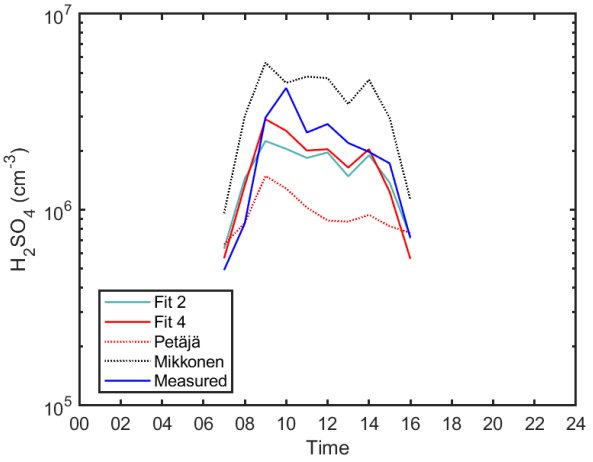
We agree with the reviewer that it is rather crucial to compare to Mikkonen et al. as it has been developed for several locations including a broad range of conditions. However, since our proxy includes periods that we have not considered previously (e.g., nighttime), we still think that it is an improvement over previous proxies.

We compared our proxies with Mikkonen et al. 2011 in all 4 locations, and added the diurnal Mikkonen plot to the main text (Figures 2,4, 6 and 8) while the scatter plots between measured sulphuric acid concentrations and both of Petäjä and Mikkonen proxies during daytime (GlobRad >= 50 W/m^2) in Figures S13 and S14, respectively.

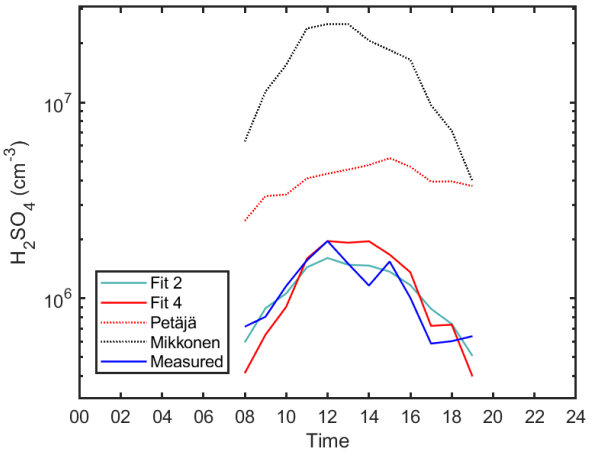
Hyytiälä



Agia Marina



Budapest



Beijing

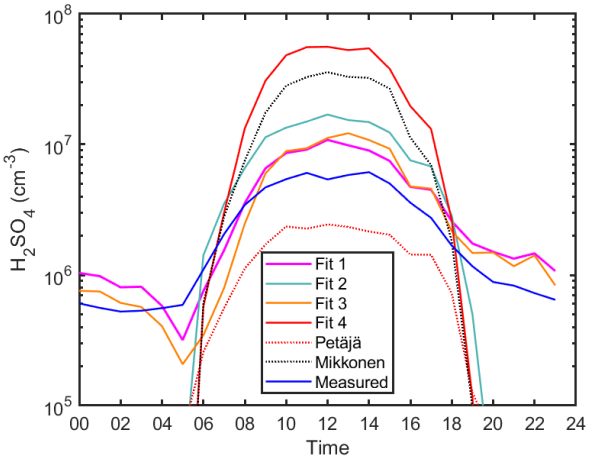
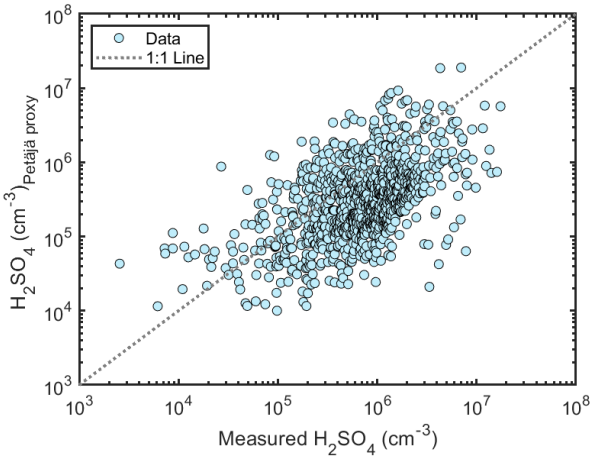
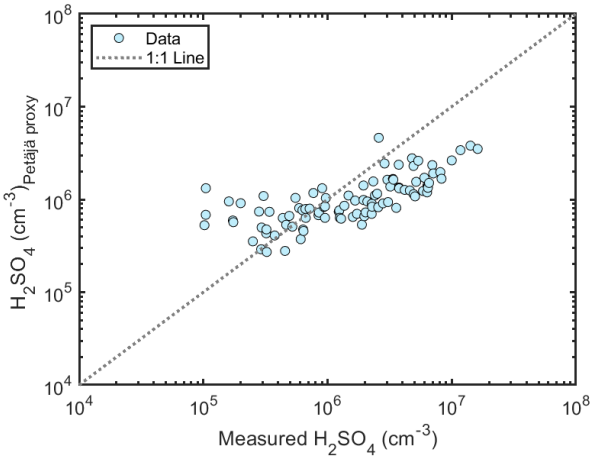


Figure R 1 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7) and Mikkonen et al. 2011 (Equation 8).

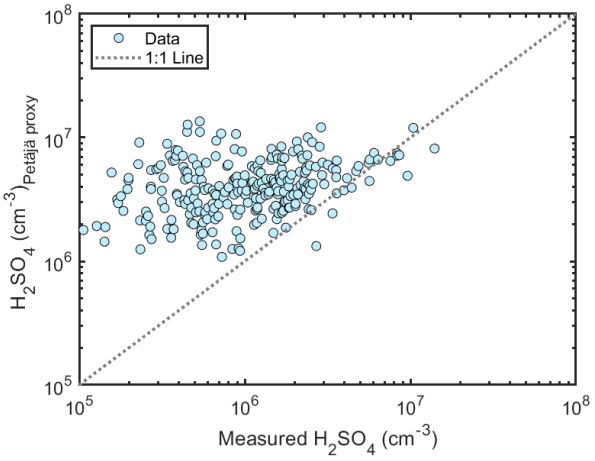
Hyytiälä



Agia Marina



Budapest



Beijing

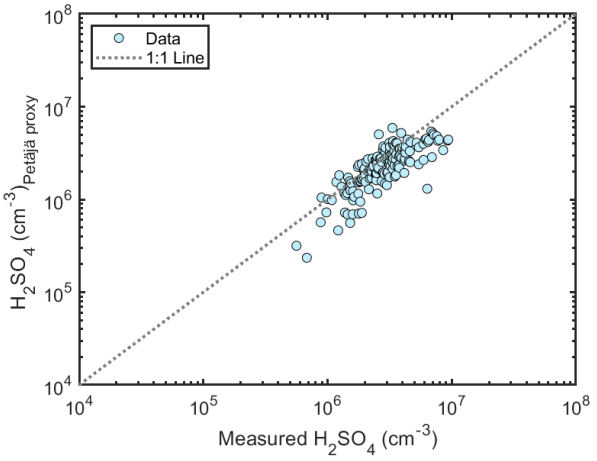
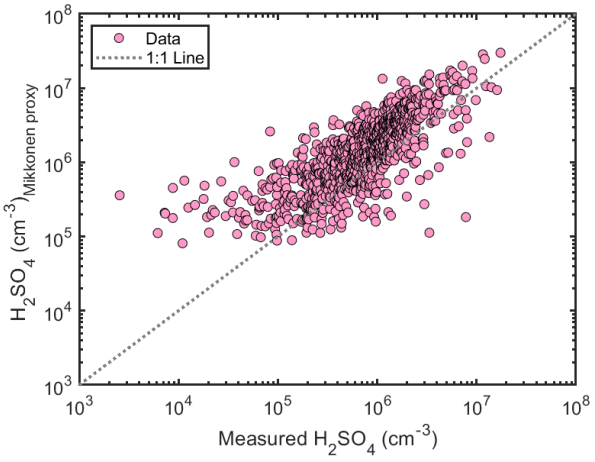
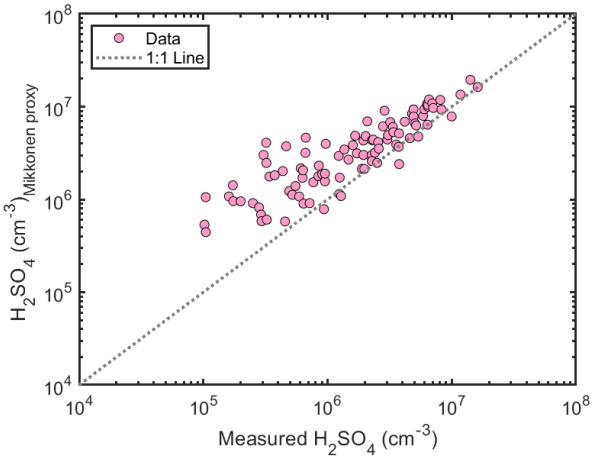


Figure R 2 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Petäjä et al. 2009 proxy at the 4 locations during daytime (GlobRad >= 50 W/m^2): Hyytiälä, Agia Marina, Budapest and Beijing.

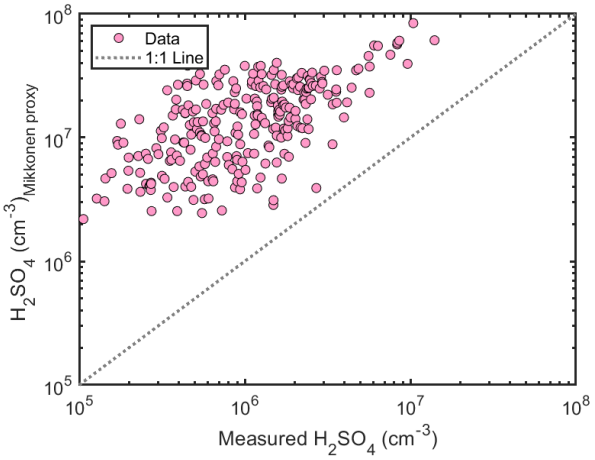
Hyytiälä



Agia Marina



Budapest



Beijing

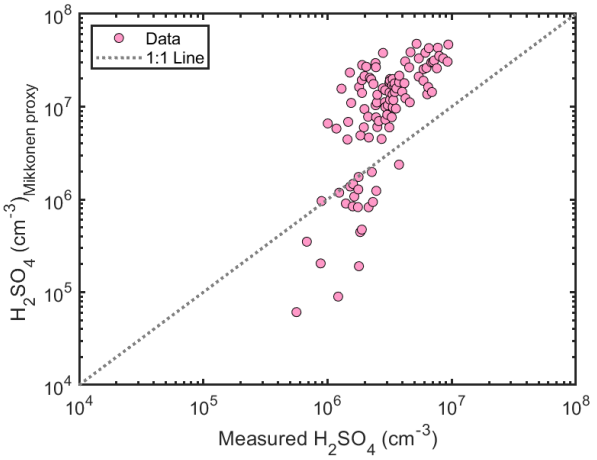


Figure R 3 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Mikkonen et al. 2011 proxy at the 4 locations during daytime (GlobRad >= 50 W/m^2): Hyytiälä, Agia Marina, Budapest and Beijing.

As minor issues:

1. Line 27: Just to be slightly fussy with wording, H2SO4 is important in new particle formation for actually two reasons: it has low volatility and also has strong intramolecular bonding abilities. Merely mentioning low volatility misses qualities that make this compound special.

We agree with the reviewer that H_2SO_4 is distinct for its strong hydrogen bonding ability which makes it possible to interact with other species and is found to be important for the first step of cluster formation. We have modified the relevant sentence on Line 58 to the following:

Sulphuric acid (H_2SO_4), which has a very low saturation vapor pressure and strong hydrogen bonding capability (Zhang et al., 2011), has been found to be the major precursor of atmospheric NPF (Weber et al., 1996; Kulmala et al., 2004; Sihto et al., 2006; Sipilä et al., 2010; Erupe et al., 2011; Lehtipalo et al., 2018; Ma et al., 2019) and is often used in global models for simulating the occurrence and intensity of new particle formation events.

2. Line 64: I suggest that the authors put a sentence or two here to state why it is important to develop a proxy for H2SO4. Many readers may be aware of the reason but it's a small thing to do and will be a great benefit to those who would otherwise be left wondering why so much effort is being placed in this.

We added the following sentences as per recommendation from the reviewer:

Line 60: *Sulphuric acid (H_2SO_4), which has a very low saturation vapor pressure, has been found to be the major precursor of atmospheric NPF (Weber et al., 1996; Kulmala et al., 2004; Sihto et al., 2006; Sipilä et al., 2010; Erupe et al., 2011; Lehtipalo et al., 2018; Ma et al., 2019) and is often used in global models for simulating the occurrence and intensity of new particle formation events (Dunne et al., 2016).*

and to Line 80:

Besides the abovementioned-previously-developed proxies, an additional proxy is still needed for representing nighttime periods which were not considered previously.

3. Line 75: I notice that Dr. Mikkonen is a reviewer of this article, so perhaps he will make this point (and I hope he also raises the concern that I express above). While the statement that his parameterization does not include condensation sink it technically correct, I believe that he considered this in his statistical analysis and found that condensation sink, or rather higher aerosol loading, is associated both with the source and sink of H2SO4, and that is the reason why on average it does not appear in the parameterization. If true then perhaps more accurate to state it this way rather than to leave the reader to conclude that this model overlooked the potential role of condensation sink.

We did not intend to say that Mikkonen et al. (2011) have overlooked the potential role of condensation sink, we have however referred to their sentence in the abstract copied below.

Sentence from Dada et al. 2020: "Proxies developed by Mikkonen et al. (2011) suggested that the sulphuric acid concentration depends mostly on the available radiation and SO_2 concentration, with little influence of CS."

Sentence from Mikkonen et al. 2011: “Interestingly, the role of the condensation sink in the proxy was only minor, since similarly accurate proxies could be constructed with global solar radiation and SO2 concentration alone.”

4. Line 86: I suggest you choose a better word than “goodness”

We modified the sentence to the following:

In order to evaluate the accuracy of the our hypothesized sources and sinks and derive the proxy equations goodness of our new proxy, we utilize measurements from four different locations: (1) Hyytiälä, Finland, (2) Agia Marina, Cyprus, (3) Budapest, Hungary and (4) Beijing, China, representing a semi-pristine boreal forest environment, rural environment in the Mediterranean area, urban environment and heavily polluted megacity, respectively. To evaluate the predictive power of the derived proxies, the equations are further tested on independent data sets.

5. Line 249: this reference to Petaja paper seems strange. Why wasn’t standard referencing used is referring to Equation 7 in the text (e.g., on line 245)?

We thank the reviewer for noticing; we modified the related text to the following:

We also refitted the data using the simple proxy proposed by Petäjä et al. (2009) by excluding the formation of sulphuric acid via stabilized Criegee intermediates source pathway and loss of sulphuric acid via the cluster formation pathway using Equation 6 and evaluated it by comparing to the original Petäjä et al. (2009) proxy using Equation 7.

$$\frac{d[H_2SO_4]}{dt} = k_1 GlobRad[SO_2] - CS[H_2SO_4] \qquad (1)$$

$$\frac{d[H_2SO_4]}{dt} = 1.4x\ 10^7 x\ GlobRad^{-0.7}[SO_2]\ GlobRad - CS[H_2SO_4] \qquad (2)$$

References

Dunne, E. M., Gordon, H., Kurten, A., Almeida, J., Duplissy, J., Williamson, C., Ortega, I. K., Pringle, K. J., Adamov, A., Baltensperger, U., Barmet, P., Benduhn, F., Bianchi, F., Breitenlechner, M., Clarke, A., Curtius, J., Dommen, J., Donahue, N. M., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Jokinen, T., Kangasluoma, J., Kirkby, J., Kulmala, M., Kupc, A., Lawler, M. J., Lehtipalo, K., Makhmutov, V., Mann, G., Mathot, S., Merikanto, J., Miettinen, P., Nenes, A., Onnela, A., Rap, A., Reddington, C. L. S., Riccobono, F., Richards, N. A. D., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Simon, M., Sipilaa, M., Smith, J. N., Stozkhov, Y., Tome, A., Trostl, J., Wagner, P. E., Wimmer, D., Winkler, P. M., Worsnop, D. R., and Carslaw, K. S.: Global atmospheric particle formation from CERN CLOUD measurements, Science, 354, 1119-1124, 10.1126/science.aaf2649, 2016.

Erupe, M. E., Viggiano, A. A., and Lee, S. H.: The effect of trimethylamine on atmospheric nucleation involving H2SO4, Atmos. Chem. Phys., 11, 4767-4775, 10.5194/acp-11-4767-2011, 2011.

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

Lehtipalo, K., Yan, C., Dada, L., Bianchi, F., Xiao, M., Wagner, R., Stolzenburg, D., Ahonen, L. R., Amorim, A., Baccarini, A., Bauer, P. S., Baumgartner, B., Bergen, A., Bernhammer, A.-K., Breitenlechner, M., Brilke, S., Buchholz, A., Mazon, S. B., Chen, D., Chen, X., Dias, A., Dommen, J., Draper, D. C., Duplissy, J., Ehn, M., Finkenzeller, H., Fischer, L., Frege, C., Fuchs, C., Garmash, O., Gordon, H., Hakala, J., He, X., Heikkinen, L., Heinritzi, M., Helm, J. C., Hofbauer, V., Hoyle, C. R., Jokinen, T., Kangasluoma, J., Kerminen, V.-M., Kim, C., Kirkby, J., Kontkanen, J., Kürten, A., Lawler, M. J., Mai, H., Mathot, S., Mauldin, R. L., Molteni, U., Nichman, L., Nie, W., Nieminen, T., Ojdanic, A., Onnela, A., Passananti, M., Petäjä, T., Piel, F., Pospisilova, V., Quéléver, L. L. J., Rissanen, M. P., Rose, C., Sarnela, N., Schallhart, S., Schuchmann, S., Sengupta, K., Simon, M., Sipilä, M., Tauber, C., Tomé, A., Tröstl, J., Väisänen, O., Vogel, A. L., Volkamer, R., Wagner, A. C., Wang, M., Weitz, L., Wimmer, D., Ye, P., Ylisirniö, A., Zha, Q., Carslaw, K. S., Curtius, J., Donahue, N. M., Flagan, R. C., Hansel, A., Riipinen, I., Virtanen, A., Winkler, P. M., Baltensperger, U., Kulmala, M., and Worsnop, D. R.: Multicomponent new particle formation from sulfuric acid, ammonia, and biogenic vapors, Science Advances, 4, eaau5363, 10.1126/sciadv.aau5363 2018.

Ma, F., Xie, H.-B., Elm, J., Shen, J., Chen, J., and Vehkamäki, H.: Piperazine Enhancing Sulfuric Acid-Based New Particle Formation: Implications for the Atmospheric Fate of Piperazine, Environ Sci Technol, 53, 8785-8795, 10.1021/acs.est.9b02117, 2019.

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petaja, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M., and Laaksonen, A.: A statistical proxy for sulphuric acid concentration, Atmos Chem Phys, 11, 11319-11334, 10.5194/acp-11-11319-2011, 2011.

Sihto, S. L., Kulmala, M., Kerminen, V. M., Dal Maso, M., Petaja, T., Riipinen, I., Korhonen, H., Arnold, F., Janson, R., Boy, M., Laaksonen, A., and Lehtinen, K. E. J.: Atmospheric sulphuric acid and aerosol formation: implications from atmospheric measurements for nucleation and early growth mechanisms, Atmos Chem Phys, 6, 4079-4091, DOI 10.5194/acp-6-4079-2006, 2006.

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, 10.1126/science.1180315 2010.

Weber, R. J., Marti, J. J., McMurry, P. H., Eisele, F. L., Tanner, D. J., and Jefferson, A.: MEASURED ATMOSPHERIC NEW PARTICLE FORMATION RATES: IMPLICATIONS FOR NUCLEATION MECHANISMS, Chemical Engineering Communications, 151, 53-64, 10.1080/00986449608936541, 1996.

Zhang, R., Khalizov, A., Wang, L., Hu, M., and Xu, W.: Nucleation and growth of nanoparticles in the atmosphere, Chem. Rev., 112, 1957-2011, 2011.

Sources and sinks driving sulphuric acid concentrations in contrasting environments: implications on proxy calculations

Lubna Dada[1,2], Ilona Ylivinkka[2,3], Rima Baalbaki[2], Chang Li[1], Yishuo Guo[1], Chao Yan[1,2], Lei Yao[1,2], Nina Sarnela[2], Tuija Jokinen[2], Kaspar R. Daellenbach[2], Rujing Yin[4], Chenjuan Deng[4], Biwu Chu[1,2], Tuomo Nieminen[2], Yonghong Wang[1,2], Zhuohui Lin[1], Roseline C. Thakur[2], Jenni Kontkanen[2], Dominik Stolzenburg[2], Mikko Sipilä[2], Tareq Hussein[2], Pauli Paasonen[2], Federico Bianchi[2], Imre Salma[5], Tamás Weidinger[6], Michael Pikridas[7], Jean Sciare[7], Jingkun Jiang[4], Yongchun Liu[1], Tuukka Petäjä[2], Veli-Matti Kerminen[2], Markku Kulmala[1,2]

[1] Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing, China.
[2] Institute for Atmospheric and Earth System Research INAR / Physics, Faculty of Science, University of Helsinki, Finland
[3] SMEAR II station, University of Helsinki, 35500 Korkeakoski, Finland
[4] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, 100084 Beijing
[5] Institute of Chemistry, Eötvös University, 1518 Budapest, P.O. Box 32, Hungary
[6]Department of Meteorology, Eötvös University, H-1518 Budapest, P.O. Box 32, Hungary
[7] The Cyprus Institute, Climate & Atmosphere Research Centre (CARE-C), 20 Konstantinou Kavafi Street, 2121, Nicosia, Cyprus

Correspondence to: markku.kulmala@helsinki.fi

Abstract

Sulphuric acid has been shown to be a key driver for new particle formation and subsequent growth in various environments mainly due to its low volatility. However, direct measurements of gas-phase sulphuric acid are oftentimes not available, and the current sulphuric acid proxies cannot predict for example its nighttime concentrations or result in significant discrepancies with measured values. Here, we define the sources and sinks of sulphuric acid in different environments and derive a new physical proxy for sulphuric acid to be utilized in locations and during periods when it is not measured. We used H_2SO_4 measurements from four different locations: Hyytiälä, Finland; Agia Marina, Cyprus; Budapest, Hungary; and Beijing, China, representing semi-pristine boreal forest, rural environment in the Mediterranean area, urban environment and heavily polluted megacity, respectively. The new proxy takes into account the formation of sulphuric acid from SO_2 via OH oxidation and other oxidation pathways, specifically that via stabilized Criegee Intermediates. The sulphuric acid sinks included in the proxy are its condensation sink (CS) and atmospheric clustering starting from H_2SO_4 dimer formation. Indeed, we found that the observed sulphuric acid concentration can be explained by the proposed sources and sinks with similar coefficients in the four contrasting environments where we have tested it. Thus, the new proxy is a more flexible and an important improvement ~~of~~ over previous proxies. Following the recommendations in the manuscript, a proxy for a specific location can be derived.

Keywords: sulphuric acid, proxy, boreal, rural, urban, megacity

1. Introduction

Atmospheric New Particle formation (NPF) events and their subsequent growth have been observed to take place almost everywhere in the world (Kulmala et al., 2004; Kerminen et al., 2018). Many of these observations are based on continuous measurements and some include more than a year of measurement data (Nieminen et al., 2018). The importance of NPF events on the global aerosol budget and cloud condensation nuclei formation has been well established (Spracklen et al., 2008; Merikanto et al., 2009; Spracklen et al., 2010; Kerminen et al., 2012; Gordon et al., 2017). Recently, the contribution of NPF to haze formation, which was still controversial, is being investigated in an increasing number of studies from Chinese megacities (Guo et al., 2014).

Sulphuric acid (H_2SO_4), which has a very low saturation vapor pressure and strong hydrogen bonding capability (Zhang et al., 2011), has been found to be the major precursor of atmospheric NPF (Weber et al., 1996; Kulmala et al., 2004; Sihto et al., 2006; Sipilä et al., 2010; Erupe et al., 2011; Lehtipalo et al., 2018; Ma et al., 2019) ~~.~~and is often used in global models for simulating the occurrence and intensity of new particle formation events (Dunne et al., 2016). However, atmospheric measurements of gas-phase sulphuric acid are rare, mainly due to its low concentration (10^6–10^7 molecules cm^{-3} or below) that can only be measured using state-of-the art instruments (Mikkonen et al., 2011) such as the Chemical Ionization atmospheric pressure interface time of flight spectrometer (CI-APi-ToF) (Eisele and Tanner, 1993; Jokinen et al., 2012). Therefore, a physically and chemically sound proxy is needed to estimate H_2SO_4 concentrations in various environments where NPF events are observed but H_2SO_4 concentrations are not continuously measured.

Due to its important participation in clustering and thus in the NPF process, several studies have tried to produce proxies for H_2SO_4 in order to fill gaps in data. For example, Petäjä et al. (2009) developed an approximation of gas-phase H_2SO_4 concentration in Hyytiälä, southern Finland, using its source from reactions between SO_2 and OH radicals, and its loss by condensation onto pre-existing particles (condensation sink, CS). Later, Mikkonen et al. (2011) developed H_2SO_4 proxies based on measurements at six urban, rural and forest areas in European and North American sites. Proxies developed by Mikkonen et al. (2011) suggested that the sulphuric acid concentration depends mostly on the available radiation and SO_2 concentration, with little influence by CS. However, Lu et al. (2019), who developed a daytime proxy based on measurement in Beijing China, ~~proved~~ suggested the need of taking into account the CS when approximating gaseous H_2SO_4, especially in areas where the condensational sink can be relatively high. The proxy developed by Lu et al. (2019) takes into consideration the formation pathways of H_2SO_4 via OH radicals from both the conventional photolysis of O_3 and from the photolysis of HONO, as well as, the loss of H_2SO_4 via CS. Besides the ~~abovementioned,~~ previously-developed proxies, an additional proxy is still needed for representing nighttime periods which were not considered previously.

Here, we derive a new proxy which takes into account the production of gaseous sulphuric acid from SO_2 with oxidation by OH and stabilized Criegee Intermediates (Mauldin et al., 2012) reactions, and its losses onto pre-existing aerosol particles (condensation sink) and due to molecular cluster formation. In order to evaluate ~~the accuracy of the~~ our hypothesized sources and sinks and derive the proxy equations ~~goodness of our new proxy~~, we utilize measurements from four different locations: (1) Hyytiälä, Finland, (2) Agia Marina, Cyprus, (3) Budapest, Hungary and (4) Beijing, China, representing a semi-pristine boreal forest environment, rural environment in the Mediterranean area,

urban environment and heavily polluted megacity, respectively. To evaluate the predictive power of the derived proxies, the equations are further tested on independent data sets. We further compare the coefficients of production and losses in each environment in order to understand the prevailing mechanism of the H_2SO_4 budget in each of the studied environments. As a result of this investigation, a well-defined sulphuric acid concentration can be derived for multiple areas around the world and even extended in time during times when it was not measured (such as: gap filling, forecast, prediction, estimation, etc.).

2. Measurement locations, observations and instrumentation

2.1. Locations

Semi-pristine boreal forest environment: Hyytiälä, Finland

Measurements were conducted at the SMEAR II-station (Station for Measuring Ecosystem–Atmosphere Relations), located in Hyytiälä (61.1° N, 24.17°E, 181 m a.s.l. (Hari and Kulmala, 2005)), southern Finland. Here we used measurements from August 18, 2016 to ~~April 16~~June~~,~~ 5, 2017 and from March 8, 2018 to February 28, 2019. ~~The measurements were performed at a tower 35 m above the ground level.~~ The data from 2016, 2018 and 2019 ~~is~~was used as a training data set for developing the proxy equation, while the data from 2017 ~~is~~was used for testing the predictive power of the developed proxy. A summary for all locations and instrumentation is given in Tables S1 (training data sets) and S2 (testing data sets).

Rural background site: Agia Marina, Cyprus

Measurements were conducted at the Cyprus Atmospheric Observatory (CAO) (35.03° N, 33.05° E; 532 m a.s.l.), a rural background site located close to Agia Marina Xyliatou village, between February 22 and March 3, 2018. For more details, see for example Pikridas et al. (2018). The data set from this location is used solely as a training data set.

Semi-urban site: Helsinki, Finland

Measurements were conducted at the SMEAR III-station, located in Helsinki (60.20° N, 24.96° E, 25 m a.s.l.). For more details about the location see for example Hussein et al. (2008). Here, we measured from July 1, 2019 to July 16 2019 as a testing data set.

Urban location: Budapest, Hungary

The measurements took place at the Budapest platform for Aerosol Research Training (BpART) Research Laboratory (47.47° N, 19.06° E~~N 47° 28' 30", E 19° 03' 45"~~, 115 m a.s.l.) of the Eötvös University situated on the bank of the Danube between March 21 and ~~May 2~~April 17, 2018. The site represents a well-mixed average atmosphere of the city centre Salma et al. (2016a). The data set from this location is used solely as a training data set.

Polluted megacity: Beijing, China

Here, observations performed at the west campus of Beijing University of Chemical Technology (39.94° N, 116.30° E) ~~during December 1, 2018 to January 31, 2019~~between March 15, 2019 and June 15, 2019 were used as a training data set while observations from September 8, 2019 to October 15, 2019 where used as a testing data set. The sampling took place from outside the window at the 5th floor of a university building adjacent to a busy street. For more details, see for example Lu et al. (2019); Zhou et al. (2020).

Near an oil-refinery industrial area: Kilpilahti, Finland

The measurement took place at Nyby measurement station (60.31° N, 25.50° E) between June 07 and June 29, 2012. The site is within 1.5 km close to Neste Oy. oil refinery and Kilpilahti industrial area. For more information on the site, please see Sarnela et al. (2015). The data set from this location is used solely as a testing data set.

2.2. **Instrumentation**

Trace Gases

A summary for all locations and instrumentation is given in Tables S1and S2.~~.~~ Measurements of different variables within the same location are performed at the same platform unless specified otherwise. In all ~~four~~ locations, the sulphuric acid concentrations were measured using a Chemical Ionization atmospheric pressure interface time of flight spectrometer (CI-APi-ToF) (Eisele and Tanner, 1993; Jokinen et al., 2012) with NO_3^- as a reagent ion and analyzed using a tofTools package based on MATLAB software (Junninen et al., 2010). In all locations, the CI-APi-ToF instruments were calibrated in a similar way prior to the campaign using the method presented by Kurten et al. (2012) to ensure the results from different sites are comparable. In Hyytiälä, the sulphuric acid concentrations were measured at the tower 35 m above ground level. In Helsinki, the sulphuric acid concentrations were measured from the 4th floor window (~12 m above ground level) of the university building adjacent (~200 m) to the SMEAR III station. In Hyytiälä, ~~-~~and Beijing, the SO_2 and O_3 concentrations were measured using an SO_2 analyzer (Model 43i, Thermo, USA), with a detection limit of 0.1 ppbv, and O_3 analyzer (Model 49i, Thermo, USA), respectively. In Hyytiälä, the trace gases concentrations were measured at the tower 16.8 m above ground level. In Helsinki, the SO_2 concentrations were monitored at a 32 m tower at the SMEAR III station using UV-fluorescence (Horiba APSA 360). In ~~Cyprus~~Agia Marina, SO_2 and O_3 are monitored using Ecotech Instruments (9850 and 9810, respectively). Concentrations of SO_2 in Budapest were measured by UV fluorescence (Ysselbach 43C) with a time resolution of 1 h at a station of the National Air Quality Network located 1.7 km in the upwind prevailing direction from the BpART site. It was shown earlier that the hourly average SO_2 concentrations (See Figure S1) in central Budapest are ordinarily distributed without large~~r~~ spatial gradients (Salma and Németh, 2019; Mikkonen et al., 2020). In Kilpilahti, SO_2 concentration were measured using Thermo Scientific ™ Model 43i SO_2 Analyser at Neste Oil refinery. Trace gases measured during the short campaign periods in Agia Marina, and Budapest are representative of yearly concentrations in respective locations when compared to longer term measurements at the same site (Salma et al., 2016b; Baalbaki, 2020, In Prep.; Mikkonen et al., 2020).

Particle number Size Distribution

The condensation sink (CS) was calculated using the method proposed by Kulmala et al. (2012) from number size distribution measurements. In Hyytiälä, the particle number size distribution was measured using a twin differential mobility particle sizer (DMPS) (Aalto et al., 2001). ~~Hygroscopic growth correction was included when calculating the CS in Hyytiälä (Figure S2).~~ In Agia Marina, the particle number size distribution between 2 and 800 nm was reconstructed from two instruments: an Airel NAIS (Neutral cluster and Air Ion Spectrometer, 2-20 nm) and TSI SMPS (Scanning Mobility Particle Sizer, 20-800 nm). In Helsinki, a twin-DMPS system (diameter 3–950 nm) was used to monitor the particle number size distribution. In Budapest, the particle number size distribution was measured by a flow-switching type DMPS in a diameter range from 6 to 1000 nm in the dry state of particles (RH<30%) in 30 channels with a time resolution of 8 min (Salma et al., 2016a). In Beijing, the particle number size distribution between 3 nm and 850 nm was measured using a Particle Size Distribution System (PSD,(Liu et al., 2016)). Condensation sink obtained at Kilpilahti was acquired from particle number size distribution measured using a DMPS (6- 1000 nm). Although having a diurnal cycle, condensation sink values obtained during the short campaign periods in Agia Marina and Budapest are representative of yearly concentrations in respective locations when compared to longer term measurements at the same site (Salma et al., 2016b; Baalbaki, 2020, In Prep.).

Radiation

In Hyytiälä, Global radiation (GlobRad) was measured using a SK08 solar pyranometer until August 24, 2017 and after that using a EQ08-S solar pyranometer. The measurements were relocated from 18-m height to 37-m height on February 14, 2017. Global Radiation from the Agia Marina is monitored using a weather station (Campbell Scientific Europe). In Helsinki, the global radiation is measured using Kipp and Zonen CNR1 at 31 m above ground level in the SMEAR III station. In Budapest, global radiation was measured by an SMP3 pyranometer (Kipp and Zonnen, The Netherlands) on the roof of the building complex with a time resolution of 1 min. Its operation was checked by comparing the measured data with those obtained from regular radiation measurements performed by a CMP11 pyranometer (Kipp and Zonnen, The Netherlands) at the Hungarian Meteorological Service (HMS) -at a distance of 10 km. The annual mean GlobRad ratio and SD of the 1-h values for the BpART and HMS stations were 1.03±0.23 for GlobRad > 100 W m^{-2}, which changed to 1.01±0.05 when considering ~~additionally~~ clear sky conditions. In Beijing, GlobRad intensity from 285 nm to 2800 nm was measured at the rooftop of the 5-floor building using a CMP11 pyranometer (Kipp and Zonnen, Delft, The Netherlands). The radiometer was maintained weekly to ensure the location horizontally and clean. In order to do the fitting for the nighttime data, zero values were replaced by the detection limit of the instrument assumed to be half the minimum measured radiation. In Kilpilahti, no global radiation measurements were available, so we relied on radiation data measured at the SMEAR III station which is around 32 km from the measurement site.~~In Beijing, GlobRad intensity was measured at the rooftop of the 5-floor building using a Vaisala Weather station data acquisition system (AWS310, PWD22, CL51), Meteon.~~

Alkenes

Volatile organic compounds (VOCs) were measured with a proton transfer reaction quadrupole mass spectrometer (PTR-MS, Ionicon Analytik GmbH) in Hyytiälä. Ambient mixing ratios are measured

every third hour from several different measurement heights. In this study, we use monoterpene ~~data~~ concentration from 16.8 m height. The instrument is calibrated regularly with standard gas (Apel-Riemer Environmental, Inc.) (Taipale et al., 2008). The same instrumentation was used to measure monoterpene concentrations in Kilpilahti every 1 hour.

In Beijing, VOCs were measured using single photon ionization time-of-flight mass spectrometer (SPI-MS 3000R, Hexin Mass Spectrometry) with unit mass resolution (UMR) (Gao et al., 2013) from September 27, 2018 to May 28, 2019. The alkenes included here are ~~propylene,~~ butylene, butadiene, isoprene, pentene and hexene. As the instrument cannot distinguish conformers, the pentene and hexene could also be cyclopentene and cyclohexene. Correlation coefficients between the different variables used in our study in all four locations are shown in Figures S2-S6.

Meteorological parameters

Temperature (T) and relative humidity (RH) in Hyytiälä were measured at 16.8 m using a 4-wire PT-100 sensors, and relative humidity sensors (Rotronic Hygromet MP102H with Hygroclip HC2-S3, Rotronic AG, Bassersdorf, Switzerland), respectively. In Agia Marina, T and RH were measured using a weather station (Campbell Scientific Europe). T and RH were measured at the Physicum rooftop 26 m above ground level and 220 m northeast from SMEAR III using a Pentronics PT100 sensor and Vaisala HMP243 transmitter, respectively. In Budapest, T and RH were measured using a Vaisala HMP45D humidity and temperature probe, at the Hungarian Meteorological Service (HMS) within a 10 km radius from the BpArt station. In Beijing, meteorological parameters are monitored by a Vaisala Weather station data acquisition system (AWS310).

3. Derivation of the new proxy

We applied the following equation to describe the time-evolution of gas-phase sulphuric acid concentration:

$$\frac{d[H_2SO_4]}{dt} = k_0[OH][SO_2] + k_2[O_3][Alkene][SO_2] - CS[H_2SO_4] - k_3[H_2SO_4]^2 \qquad (1)$$

Here, k_o represents the coefficient of H_2SO_4 production term due to the well-known SO_2 - OH reaction (Petäjä et al., 2009) and k_2 is the coefficient of H_2SO_4 production via stabilized Criegee Intermediates (sCI) produced by the ozonolysis of alkenes (Mauldin et al., 2012). Here we use available monoterpene concentration (MT) as a proxy for alkenes in Hyytiälä as they are the dominating species in the boreal forest environment (Hakola et al., 2012; Hellén et al., 2018; Rinne et al., 2005). For Beijing, we use urban dominating aromatic alkenes. As no VOC measurements are performed in neither Agia Marina nor Budapest, we evaluate the proxy without the stabilized Criegee Intermediate source term. It is important to note here that the coefficient for sCI is a "bulk" term, and it varies from place to place due to the differences in sCI structures and different production efficiency from different alkene species (Novelli et al., 2017; Sipilä et al., 2014). The third term in Equation 1 represents the loss of H_2SO_4 onto pre-existing aerosol particles, known as condensation sink (CS) and is calculated by multiplying the CS calculated for sulphuric acid with the concentration of sulphuric acid monomer. The fourth term in Equation 1 is defined as the square of sulphuric acid concentration multiplied by clustering coefficient k_3. The square of sulphuric acid represents the collision of two sulphuric acid monomers forming a sulphuric acid dimer which was found to be the first step of

atmospheric cluster formation (Yao et al., 2018). Therefore, this term takes into account the additional loss of H_2SO_4 due to cluster formation not included in the term containing CS. This is necessary because CS is only inferred from size-distribution measurements at maximum down to 1.5 nm, i.e. not containing any cluster concentrations and hence losses onto these clusters. This term is written in the form of sulphuric acid dimer production, which seems to be the first step of cluster formation once stabilized by bases (Kulmala et al., 2013; Almeida et al., 2013; Yao et al., 2018).

Since measuring the OH concentration is challenging, we first replaced it with the UVB radiation intensity, which has been shown to be a good proxy for the OH concentration (Berresheim et al., 2002; Lu et al., 2019; Rohrer and Berresheim, 2006). Unfortunately, UVB was not measured in all the field studies considered here. Alternatively, GlobRad, a commonly measured quantity, tends to correlate well with UVB and can generally replace it, as used previously by Petäjä et al. (2009). We confirmed the strong correlation between UVB radiation and Global radiation in two locations, Hyytiälä and Beijing (Figure S7-S8). Accordingly, the coefficient k_1 here replaces the coefficient of H_2SO_4 production k_o terms (Equation 2). We proceed here using only GlobRad in the proxy to be consistent with the two other locations where UVB was not measured (Agia Marina and Budapest).

$$\frac{d[H_2SO_4]}{dt} = k_1 GlobRad[SO_2] + k_2[O_3][Alkene][SO_2] - CS[H_2SO_4] - k_3[H_2SO_4]^2 \quad (2)$$

By assuming a steady state between H_2SO_4 production and loss, the H_2SO_4 concentration can be solved directly from Equation (2):

$$[H_2SO_4] = -\frac{CS}{2k_3} + \left[\left(\frac{CS}{2k_3}\right)^2 + \frac{[SO_2]}{k_3}(k_1 GlobRad + k_2[O_3][Alkene])\right]^{\frac{1}{2}} \quad (3)$$

~~$[H_2SO_4] = -\frac{CS}{2k_3} + \sqrt{\left(\frac{CS}{2k_3}\right)^2 + \frac{[SO_2]}{k_3}(k_1 GlobRad + k_2[O_3][Alkene])}$ (3)~~

In order to evaluate the importance of each of the source terms in determining the change in sulphuric acid concentration, we refitted the data after excluding the stabilized Criegee intermediates source pathway as shown in Equation 4.

$$\frac{d[H_2SO_4]}{dt} = k_1 GlobRad[SO_2] - CS[H_2SO_4] - k_3[H_2SO_4]^2 \quad (4)$$

In order to evaluate the importance of each of the sink terms in determining the sulphuric acid concentration, we refitted the data after excluding the loss of sulphuric acid via the cluster formation pathway using Equation 5.

$$\frac{d[H_2SO_4]}{dt} = k_1 GlobRad[SO_2] + k_2[O_3][Alkene][SO_2] - CS[H_2SO_4] \quad (5)$$

we also refitted the data using the simple proxy proposed by Petäjä et al. (2009) by excluding the formation of sulphuric acid via stabilized Criegee intermediates source pathway and loss of sulphuric

acid via the cluster formation pathway using Equation 6 and evaluated it by comparing to the original Petäjä et al. (2009) proxy using Equation 7 and Mikkonen et al. (2011) using Equation 8. The calculation of the scaled reaction constant *k* used in Equation 8 is given in the supplementary material section 1.

$$\frac{d[H_2SO_4]}{dt} = k_1 GlobRad[SO_2] - CS[H_2SO_4] \quad (6)$$

$$\frac{d[H_2SO_4]}{dt} = 1.4x\ 10^{-7} x\ GlobRad^{-0.7}[SO_2][GlobRad] - CS[H_2SO_4] \quad \text{~~(See Petäjä et al. 2009)~~} \quad (7)$$

$$[H_2SO_4] = 8.21\ x\ 10^{-3}\ k\ GlobRad[SO_2]^{0.62}(CS.RH)^{-0.13} \quad (8)$$

The equations derived for each of the sites can be found in Table 1. The fitting coefficients were obtained by minimizing the sum of the squared logarithm of the ratio between the proxy values and measured sulphuric acid concentration using the method described by Lagarias et al. (1998), a build-in function *fminsearch* of MATLAB, giving the optimal values for the coefficients. The data were~~was~~ subject to 10,000 boot-strap resamples~~ping~~ when getting each of the *k* values as a measure of accuracy in terms of bias, variance, confidence intervals, or prediction error (Efron and Tibshirani, 1994). We accounted for the systematic uncertainty in H_2SO_4 and predictor variables. For every bootstrap fit, we assumed both H_2SO_4 and all predictor variables to be affected by independent systematic errors between its lower and upper accuracy limits. More details on the bootstrap resampling method and uncertainty introduction can be found in the supplementary information. The ~~median,~~ 25th percentile and 75th percentiles of the coefficients are shown ~~in~~ for all locations together with the median *k* values in Table ~~1~~2. The median *k* values from the bootstrap resamples were used in the equations for deriving sulphuric acid concentrations at each site. Figures S2-S6 present the correlation matrix between the different variables participating in H_2SO_4 formation and loss in all locations. In Beijing, the Alkenes (AVOCs) have different patterns in day and night which forces us to have two separate equations for daytime and nighttime. The goodness of the fit and the probability of overfitting or under-fitting was evaluated using the Akaike information criterion (Figure S9), which also compares the proxies given in equations 2, 4, 5 and 6. The criterion uses the sample size (number of points), the number of parameters (terms in the equation) and the sum of squared estimate of errors (SSE: deviations predicted from actual empirical values of data) to estimate the quality of each model, relative to each of the other models and thus provides means for model selection (McElreath, 2018).

4. Results and Discussions

4.1. The sulphuric acid proxy for Hyytiälä SMEAR II station

Figure 1 shows the scatter plot between the observed H_2SO_4 concentrations and that derived by the proxy using the full Equation 2. The correlation coefficient was 0.~~85~~ 84 (~~2089~~ 1860 data points). The data were related to 3-hour medians, as the monoterpene concentration was measured only every third hour. In Figure 1B-D, the proxy is refitted after removing one of the source or sink terms (Equations 4-6), in order to evaluate the sensitivity of the proxy to each of the terms and to show the improvement of the proxy using the additional source and sink (Figure 1A) in comparison to the simple proxy that was used by Petäjä et al. (2009) (Figure 1D). Our results show that the integration of additional terms of H_2SO_4 formation (i.e. the stabilized Criegee Intermediates) and loss (atmospheric cluster formation) gives the new proxy the ability to accurately capture the diurnal variation of the H_2SO_4 concentration, demonstrating a clear improvement over the earlier physical proxy (Petäjä et al., 2009). In Figure 1B the corresponding data are shown without the alkene term (Equation 4). The correlation is ~~significantly~~ substantially weaker (0.~~73~~70) than with the full equation. Even more importantly, we cannot estimate the contribution of the alkene term to the sulphuric acid concentration (Figure 2 – Fit 2) as the fit results also in an unphysical coefficient for cluster formation (Kürten et al., 2015) and the fit fails to capture the diurnal pattern during dark hours after 16:00 (Figure 2 – Fit 2). When fitting the data without the cluster source term (Equation 5), the correlation coefficient is high (Figure 1C), yet the goodness of the fit is not as good as when the cluster source term is taken into account (Table S4 - Figure S9). Furthermore, we derived an additional proxy equation using CS corrected for hygroscopic growth (Laakso et al., 2004) to be used when calculating a more robust proxy for Hyytiälä. The details, equation and results are shown in the supplementary information (Figure S10-S12).

Note that we opted for deriving a bulk proxy (daytime and nighttime together) instead of two independent proxies, one for daytime and one for nighttime separately. Our results show that one bulk equation is able to explain the Hyytiälä sulphuric acid daytime and nighttime sources accurately. Additionally, separating the bulk equation into two distinct equations results in bias towards the pattern of one of the predictor variables. For instance, the k_1 value during daytime follows the cycle of global radiation, while that of k_2 follows the cycle of alkenes. Therefore, in order to accurately reflect the continuum of source and sink terms throughout the day, we decided on the bulk proxy. Additionally, one bulk equation was able to predict sulphuric acid concentrations during daytime and nighttime with high accuracy (slope of ~1) as further discussed in section 4.5.

The fit was able to reproduce the sulphuric acid concentration in such clean environment without the cluster term (Figure 2 – Fit 3), perhaps due to low concentrations of bases participating in clustering in Hyytiälä (Jen et al., 2014). Finally, the corresponding data without both the alkene source term and cluster formation source term (Equation 6, Figure 1D) shows a weaker correlation between the measured and modelled sulphuric acid concentration (0.~~73~~70), but more importantly, it deviates far from the 1:1 line during both daytime and nighttime (Figure 2 – Fit 4). It is important to note here that when deriving the Petäjä proxy (Petäjä et al. 2009), the model relied on summer data between April and June 2007 which could explain the misfit with the current data from Hyytiälä which spans the whole year. See also figures S13 and S14 for scatter plots comparing the measured sulphuric acid concentrations of the training data set with Petäjä et al. 2009 and Mikkonen et al. 2011, respectively. In general, using all four terms in equation 2 shows improvement over all other combinations

(Equations 4-6) in terms of not only correlation coefficients and accurate diurnal cycle between measured and calculated concentrations of sulphuric acid as shown in Figures 1 and 2, but also show a better goodness of the fit as shown in Table S4 and Figure S9 when using the AIC statistical method. The final equation for the boreal forest environment can be found in Table 1, Equation 9.
~~1012~~

4.2. **Sulphuric Acid Proxy at a Rural Site: Agia Marina, Cyprus**

Since there were no direct measurements of alkenes in Agia Marina, we had to exclude the formation of H_2SO_4 in the oxidation by sCI from the proxy, and therefore we derived only the daytime H_2SO_4 proxy concentration. The correlation between the measured and proxy concentration of H_2SO_4 was 0.88 (96 data points) which shows that the chosen predictors were able to explain the measured sulphuric acid concentration largely ~~proves the truthfulness of this proxy~~ (Figure 3). However, the slope deviates from the 1-to-1 line which could be attributed to the additional formation mechanisms that we could not include with the current data. However, the addition of the cluster loss mechanism shows a noticeable improvement over the simple proxy, in Figure 3B (R = 0.80). The cluster loss term starts to become more important in this rural environment in comparison to the boreal forest, which could be due to a higher concentration of stabilizing bases in Agia Marina compared with Hyytiälä. Although both fits of, Equation 4 and 6, show similar diurnal patterns (Figure 4, Fits 2 and 4), the loss term due to H_2SO_4 cluster formation improved the precision of the new proxy (Figures 3). According to the statistical AIC method, the goodness of the fit has improved from ~~161~~ 70 to 33~~71~~, with and without the clustering term, respectively, as shown in Figure ~~S10~~S9. Also, even without the alkene term, the newly derived coefficients improved the proxy in comparison to Petäjä et al. (2009) and Mikkonen et al. (2011) as shown in Figures 4, S13 and S14. The final equation for the rural site can be found in Table 1, Equation 10.

4.3. **Proxy for urban environment: Budapest, Hungary**

Next we try to understand the mechanisms of sulphuric acid formation and losses in an even more complex environment, such as urban Budapest (Figures 5 & 6). Since there were no direct measurements of alkenes there, neither its proxies such as monoterpenes or anthropogenic volatile organic compounds, we derived the sulphuric acid proxy excluding the formation due to stabilized Criegee Intermediate pathway, as in Equation 4. In comparison to the simple proxy (Figure 5B; R = 0.49; ~~262~~ 263 data points), the correlation between the measured and proxy concentration of H_2SO_4 improved with the addition of the loss term due to cluster formation, R = 0.59 (Figure 5A). The correlation between measured and modelled values of sulphuric acid became weaker in Budapest in comparison to Hyytiälä and Agia Marina, which could be attributed to a more complex environment, and additional pathways of sulphuric acid formation and losses. Additionally, we observed a sudden SO_2 concentration change in the middle of the campaign, possibly due to sudden change in local meteorology and airmass transport, which could also explain the weaker correlation (See Figure S1). The loss term due to H_2SO_4 dimerization improved the precision of the new proxy in comparison to the simple model as well as the Petäjä et al. (2009) or the Mikkonen et al. (2011) derivation, as shown in Figure 6, S13 and S14). We think that the overestimation in the Petäjä proxy is because of its dependence on the SO_2/CS ratio. The proxy is originally derived in Hyytiälä and when we apply the same coefficients to Budapest it gives higher estimated concentration compared to the measured since SO_2/CS ratio is smaller in Budapest (Figure ~~10~~9). Although the proxy developed by Mikkonen et al.

(2011) has shown to work in varying environments, it clearly overestimates the sulphuric acid concentration in Budapest for perhaps the same reasons (its dependence on the SO_2/CS ratio). It is also visible from Figures 5 and 6, that the addition of the dimerization term was capable of better capturing the lower H_2SO_4 concentrations in comparison to fitting the data without the dimerization term. In comparison to both Hyytiälä and Agia Marina, the coefficient associated with dimerization in Budapest is slightly higher, which can be attributed to the availability of a possibly facilitated clustering due to higher abundance of stabilizing bases such as amines and ammonia (discussed in section 4.6 ~~4.5~~). The final equation for the urban environment can be found in Table 1, Equation 11.

4.4. **Proxy for Megacity: Beijing, China**

In megacities, in our case Beijing, the sulphuric acid concentration is particularly high during nighttime, which confirms the need for determining the contribution of sources other than OH (radiation) to its formation. Our observations emphasize the contribution of the alkene pathway, as without considering this route we would not replicate morning hours correctly. During daytime, there is enhanced dimerization and cluster formation due to the abundance of stabilizing bases (Yao et al., 2018). We assessed the derivation of the proxy equation first using daytime data and nighttime data separately, and found that such a separation results in an unphysical k_3 value since clustering in Beijing happens mostly during daytime(Zhou et al., 2020). This obstacle was also observed when deriving a bulk equation. To overcome it, we set an upper limit for the k_3 value at 7 x 10^{-9} obtained from the fitting of daytime data (GlobRad >= 50 W/m^2). The reason for such an observation is that~~Besides~~, in such a complex environment, sulphuric acid might originate from sources other than the ones we accounted for in our calculation especially during nighttime, for example through the hydrolysis of SO_3 formed from non-photochemical processes (Yao et al., 2020, In Rev.). ~~As a result, we derived two separate sets of equations, as shown in Table 1. Results of a combined equation are shown in Figures S11 and S12. In addition,~~The alkenes or volatile organic compounds during daytime are different from those during nighttime, and might vary between seasons, which could be attributed to a different fleet composition during those times or the biogenic activity (Yang et al., 2019). However, the derived equation 12 (derived from spring data) is able to predict the daytime and nighttime sulphuric acid concentrations during summer and autumn (See more in section 4.5) ~~. For that purpose, we had to divide the data for Beijing into two groups: daytime (GlobRad >= 50 W/m^2) and nighttime (GlobRad<50 W/m^2).~~

In Figure 7, we see an improvement of the new proxy (Equation 2) in comparison to the simple proxy (Equation 6) derived by Petäjä et al. (2009) as the former takes into the account the additional sources and sinks of H_2SO_4 which were not considered in previous works (See also Figure ~~S10~~S9). Introducing the alkene production term improved the accuracy of the H_2SO_4 concentration ~~slightly for daytime and significantly~~ during both daytime and nighttime (Figures 7 and 8), which supports our assumption that H_2SO_4 formation during nighttime is driven by stabilized Criegee Intermediates. In Figure 7B we show the proxy without the alkene term~~. Although the correlation improves, this is only because the nighttime values are not captured~~ is unable to capture the nighttime concentrations. In Figure 9, we see the importance of all sources and sinks predicted for sulphuric acid, as Fit 1 (Equation 2) predicts best the measured sulphuric acid concentration. Additionally, according to the statistical AIC method, using the full equation has the least probability of inaccuracy and error in estimating the sulphuric acid concentration (Figure ~~S10~~S9). Moreover, it is clear that the addition of the cluster sink term in Megacity environment is required due to its large contribution as a sink for

H_2SO_4 especially due to higher concentrations of stabilizing molecules, the cluster mode (sub-3 nm) particle concentration, are the highest in Chinese Megacities (Zhou et al., 2020). The final equation for the megacity can be found in Table 1, Equation 12.

4.5. **Predictive power of proxy equations**

Each of the proxies of the boreal forest environment, rural background and megacity were tested for predictive power on independent data sets using extended data sets from the same location or using measurements from locations with similar characteristics. The sulphuric acid concentrations at each of these locations is modelled using the equation (with median k per source/sink term) relevant to the site and compared to the measured concentrations. The derivation of the sulphuric acid concentrations using 10,000 combinations of k values as well as the error on the predictions are shown in the supplementary information. Note that the testing data sets are not subject to any boot strap resampling or uncertainty additions, but are rather used as is for testing the predictive power of the suggested proxy.

4.5.1 Boreal forest environment: Hyytiälä

For testing the predictive power of the boreal forest proxy (Equation 9), we use an independent testing data set from the same location measured from January 1, 2017 to June 5, 2017. Results show that the modelled sulphuric acid concentrations correlate well (R = 0.7) with the measured sulphuric concentrations with a slope of 0.997 for the testing data set (Figure 10A and S16). Moreover, we tested the four fits on the testing data set; i.e. the full Equation 2, the equation without the Stabilized Criegee Intermediates source (Equation 4), the equation without the cluster sink term (Equation 5) and the equation without neither the Stabilized Criegee Intermediates source nor the cluster sink term (Equation 6), and found that Fit 1 (Equation 4) best defines the measured sulphuric acid concentration in comparison to the rest of the equations (Figure S17). The diurnal cycle is also accurately described by the Equation 4 which captures both nighttime and daytime (Figure S18).

4.5.2. Semi-urban location: Helsinki

For testing the predictive power of the rural background site proxy (Equation 10), we use an independent testing data set from a semi-urban location in Helsinki, Finland measured from July 1, 2019 to July 16, 2019 during daytime (GlobRad >= 50 W/m^2). The rural background site equation 10 is used as the condensation sink and SO_2 concentrations in the testing location are within the interquartile span of the Agia Marina measurements (Figure 9, Table S3). Results show that although the modelled sulphuric acid concentrations do not correlate as well as in other locations (R = 0.44), the bias could be attributed to the missing source (alkene) in the original equation (Figure 10B). Indeed, looking at the binned data, we find that at within each concentration bin the modelled sulphuric concentrations tend to span the 1:1 line. Actually, the discrepancy between the measured and the modelled concentration is smaller than the model prediction error (Figure S19). Note that the model prediction error is estimated as the interquartile range of the modelled H_2SO_4 concentration of a single point in time arising from the uncertainty in k values. For the rural background site, we also found that the diurnal cycle is better described when introducing the additional clustering sink term (Figure S20).

4.5.3. Megacity: Beijing

For testing the predictive power of the megacity proxy (Equation 12), we use an independent testing data set from the same location (Beijing) measured from September 1, 2019 to October 15, 2019. Results show that the modelled sulphuric acid concentrations correlate well (R = 0.83) with the measured sulphuric concentrations with a slope of ~1.1 for the testing data set (Figure 10C). Also for

this site, we tested the four fits on the testing data set; i.e. the full Equation 2, the equation without the Stabilized Criegee Intermediates source (Equation 4), the equation without the cluster sink term (Equation 5) and the equation without neither the Stabilized Criegee Intermediates source nor the cluster sink term (Equation 6), and found that Fit 1 (Equation 4) best defines the measured sulphuric acid concentration in comparison to the rest of the equations (Figure S22). The diurnal cycle is also described by the Equation 4 which captures both nighttime and daytime (Figure S23).

4.5.4. Industrial area: Kilpilahti

Finally, we tested the predictive power of our developed proxy on a data set measured at an industrial area in close proximity to an oil refinery. Interestingly, the median CS at the location lies within the interquartile range of the CS measured in Hyytiälä and that measured in Agia Marina (Table S3, Figure 9). The SO_2 concentrations at the measurement site are higher than in both Hyytiälä and Agia Marina, but smaller than the ones reported in Budapest. Additionally, we observed alkene concentrations at Kilpilahti, which are within the range of those monitored in Hyytiälä attributed to the green belt in the area (Sarnela et al., 2015). Accordingly, we test the proxy equation 9 on the Kilpilahti data set. Our results show that Equation 9 is able to predict the sulphuric acid concentrations in Kilpilahti with a high correlation coefficient (R= 0.74) (Figure 10D). Similar to other locations, the Fit 1 (Equation 4) best describes the sources and sinks at the location (Figure S25). The discrepancy between the measured and the modelled concentration is smaller than the model prediction error for less than 50% of the data points only (Figure S24). This observation is consistent with the diurnal cycle (Figure S26). During certain mornings (4:00 – 8:00 LT), when the measured sulphuric concentrations are particularly high, the model was unable to predict the concentrations accurately. These high concentrations were attributed to air masses coming from the oil refinery (Sarnela et al., 2015). Indeed, our proxy was not able to explain these morning peaks using biogenic alkenes, however, in such an industrial area, anthropogenic sources could play a role in determining the magnitude of sulphuric acid concentrations. With the condensation sink being rather low (median ~0.005 s^{-1}), the impact of direct H_2SO_4 emissions cannot be ruled out either.

4.5.4.6. Sensitivity of the proxy to the H_2SO_4 sources and sinks

The variations of coefficients related to Equation 3 can be used to get insights into the general chemical behavior under current atmospheric conditions, as well as into the mechanisms of sulphuric acid formation and losses in various environments. The contribution of different terms in different locations seem to vary significantly. The new loss term taking into account clustering starting from dimer formation needs to be taken into account in all the environments in daytime. On the other hand, without alkene term it is in practice impossible to get nighttime concentrations correct.

In Table 1-2, we have presented the fitted coefficients (Equation 3) for all our sites, whereas the contributions of the different terms in the balance equation are given during daytime in Figure 9-11 and Table 2-3. The contribution of the various source and sink terms to the change of H_2SO_4 concentrations are determined using Equation 2. The median derived k_1, k_2 and k_3 values, together with the measured H_2SO_4, CS, trace gases and GlobRad per site, were used to calculate each of the terms. Source term 1 refers to k_1 x GlobRad x $[SO_2]$, source term 2 refers to k_2 x $[O_3]$ x [Alkene] x $[SO_2]$, sink term 3 refers to k_3 x $[H_2SO_4]^2$ and sink term 4 refers to CS x $[H_2SO_4]$. The contribution of each term is then calculated as the median or percentiles of the normalized term to the sum of all terms. The variability of the coefficients (Table 1-2), as well as the relative contributions of each term to the total sulphuric acid concentration (Table 2-3), could give valuable information on the mechanisms resulting in sulphuric acid formation and losses. At steady state (Equation 2), the sources and sinks are in balance with each other during both daytime and nighttime, but there were clear differences in the individual contributions. For instance, a variation in k_1 could be due to variations

in OH sources and sinks. Although in urban locations OH sinks are expected to be higher and therefore k_1 to be lower, additional sources of OH are available in such locations, for example HONO (Zhang et al., 2019). The alkene/Criegee intermediate term was found to be an important H_2SO_4 source (Figures 1, 2, 7 and 8), as without it we are not able predict night or morning concentrations of H_2SO_4 properly. The alkene source term contributed up to almost 100% of the H_2SO_4 sources during nighttime in Beijing and up to ~~82~~90% of the sources during nighttime in Hyytiälä (~~Table 2~~Figure 12). ~~The alkene term is, however, not only important during nighttime but also during daytime, as it contributed to the sources by a median of 41% during daytime in Beijing.~~The Criegee intermediate term showed its importance mostly when global radiation is low, not only in nighttime but also during winter (Figure 12) in both Hyytiälä and Beijing. It is important to note here that Criegee intermediates vary between locations, they also form in different yield percentages from different alkenes (Novelli et al., 2017; Sipilä et al., 2014). These stabilized Criegee intermediates also react differently under different environmental conditions.

The CS term had the highest contribution to the total sink in Hyytiälä. Its contribution decreased when moving towards more polluted environments (Figure 11), to become in Beijing, regardless of the relatively high condensation sink in Megacities, smaller than that of the cluster sink term (Laakso et al., 2006; Monkkonen et al., 2005; Monkkonen et al., 2004; Yao et al., 2018).~~,~~. This observation might be attributed to decreased effectiveness of condensation sink in more polluted environments (Kulmala et al., 2017), but also to increased contribution of the clustering sink term in such environments where the concentration of stabilizing bases is highest, particularly in daytime (Yao et al., 2018; Yan et al., 2018). It should be noted that measurements of ammonia and similar bases are rare, so their exact contribution is difficult to estimate. The cluster term is found to contribute most during spring daytime in Hyytiälä (Figure 12 – A & C), which is the time window during which clustering and thus new particle formation events happen (Dada et al., 2018; Dada et al., 2017).~~.~~ The same is observed for Beijing, where the clustering term contributed up to 70% of the total sink terms during daytime (Figure 12-D) especially during summer when the CS is lowest (Deng et al., 2020).

5. Conclusions and recommendations

Sulphuric acid is a key gas-phase compound linked to secondary aerosol production in the atmosphere. The concentration of sulphuric acid in the gas phase is governed by source and sink terms. In this paper we define the sources and sinks of H_2SO_4 and derived a physically and chemically sound proxy for the sulphuric acid concentration using measurements at 4 different locations, including boreal forest environment (Hyytiälä, Finland), a rural Mediterranean site (Cyprus), an urban area (Budapest) and a megacity (Beijing). When describing the change in gas phase sulphuric acid concentration, we took into account two source terms: 1) photochemical oxidation of sulfur dioxide and 2) sulphuric acid originating from alkene and ozone reactions and associated stabilized Criegee radical pathway. For the sink terms, we considered 3) the loss rate to the pre-existing aerosol described by condensation sink, and 4) loss rate of sulphuric acid monomer due to clustering process.

In general, the variation in the environmental conditions and difference in concentrations of air pollutants affects the coefficients derived and therefore it is important to derive location specific coefficients. The derived coefficients give insights into the general chemical behavior and into the mechanisms of sulphuric acid formation and losses in various environments. As improvements from previously derived proxies, without the alkene H_2SO_4 formation pathway, it is in practice impossible to get nighttime concentrations. On the other hand, the additional loss term taking into account

clustering starting from dimer formation needs to be taken into account in all the environments especially those with higher cluster formation probabilities due to availability of stabilizing bases.

The coefficients derived do not differ substantially between the different locations. The proxy could therefore be used at locations with no prior H_2SO_4 measurements, provided that the environmental conditions are approximately similar to those in one of the four sites described here. More specifically, the proxies could be utilized to derive long-term data sets for H_2SO_4 concentrations, which would be essential in performing various kinds of trend analyses. In order to derive the long term sulphuric acid concentrations, we recommend deriving in-house coefficients in case sulphuric acid concentrations are directly measured rather than using the ones from already derived studies. The choice of equation depends on the availability of the data on site. In case alkenes or their proxies are measured and sulphuric acid is measured, derivation of the coefficients should be based on Equation 2. In case neither alkenes nor their proxies are measured but sulphuric acid is measured, the coefficients and therefore the proxy for daytime only can be derived, using Equation 4. In case, sulphuric acid is not measured, one can calculate the sulphuric acid proxy using the Equation 2 or Equation 4, depending on whether the alkene data is available or not, respectively, using the coefficients suggested in Table 1 which are relevant to the site of interest. In order to make the best choice for the coefficients, Figure ~~10~~ 9 can be followed in order to decide which description fits the location of interest best. For instance, in case the condensation sink is between 2 x 10^{-3} and 6 x 10^{-3} s^{-1}, and the SO_2 concentration is lower than 2 x 10^9 molecules. cm^{-3}, coefficients of Hyytiälä or the boreal forest are to be used.

Data availability

The data used in the manuscript and the MATLAB code which provides the k values are available from the first author at lubna.dada@helsinki.fi.

Author contributions

MK came up with the idea, LD, IY, CL, RB analyzed the data, YG, CD, RY, CY, LY, JJ, YL, BC, ZL, YW performed the measurements in Beijing and pre-processed the raw data, NS, TJ, MS, TP performed the measurements in Hyytiälä and pre-processed the raw data, LD, TN, JK, KRD, DS, TH, PP, FB, VMK, MK provided useful discussion and ideas, IS, TW, RB, TJ performed the measurements in Budapest and pre-processed the raw data, MP, JS, RB, TJ performed the measurement in Agia Marina and pre-processed the raw data. RCT, TJ, MS performed the sulphuric acid measurements in Helsinki and pre-processed the raw-data. LD and KRD introduced the error and bootstrap resampling analyses. LD, VMK and MK wrote the manuscript. All co-authors contributed to reviewing the manuscript and to the discussions related to it.

Competing interests

All authors declare no competing interests.

Acknowledgements

This project has received funding from the ERC advanced grant No. 742206, ERC-StG No. 714621, the Academy of Finland Center of Excellence project No. 307331, Academy of Finland project No. 316114 and 296628, the National Natural Science Foundation of China project No. 41877306 and from National Key R&D Program of China (2017YFC0209503). This project receives funding from

the European Union's Horizon 2020 research and innovation program under grant agreements (ACTRIS) No. 654109 and 739530. Funding by the National Research, Development and Innovation Office, Hungary (K116788 and K132254) is acknowledged. We thank V. Varga and Z. Németh of the Eötvös University for their help in the experimental work in Budapest, ~~and~~ K. Neitola and T. Laurila for their help at Agia Marina, LJJ. Quéléver and T. Lehmusjärvi for their help in setting up the sulphuric acid measurement in Helsinki. This publication has been produced within the framework of the EMME-CARE project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 856612 and the Cyprus Government. The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Tables and Figures

Table 1 Equations for sulphuric acid proxy derivation at each of the measurement locations.

$$[H_2SO_4]_{boreal} = -\frac{CS}{2\,x\,(4.2\,x\,10^{-9})} + \left[\left(\frac{CS}{2\,x\,(4.2\,x\,10^{-9})}\right)^2 + \frac{[SO_2]}{(4.2\,x\,10^{-9})}(8.6\,\text{x}\,10^{-9}\,x\,GlobRad + 6.1\,\text{x}\,10^{-29}[O_3][Alkene])\right]^{1/2} \quad (9)$$

$$[H_2SO_4]_{rural} = -\frac{CS}{2\,x\,(2.2\,x\,10^{-9})} + \left[\left(\frac{CS}{2\,x\,(2.2\,x\,10^{-9})}\right)^2 + \frac{[SO_2]}{(2.2\,x\,10^{-9})}(9.7\,\text{x}\,10^{-8}\,x\,GlobRad)\right]^{\frac{1}{2}} \quad (10)$$

$$[H_2SO_4]_{urban} = -\frac{CS}{2\,x\,(9.8\,x\,10^{-9})} + \left[\left(\frac{CS}{2\,x\,(9.8\,x\,10^{-9})}\right)^2 + \frac{[SO_2]}{(9.8\,x\,10^{-9})}(1.57\,\text{x}\,10^{-9}\,x\,GlobRad)\right]^{\frac{1}{2}} \quad (11)$$

$$[H_2SO_4]_{megacity} = -\frac{CS}{2\,x\,(7.0\,x\,10^{-9})} + \left[\left(\frac{CS}{2\,x\,(7.0\,x\,10^{-9})}\right)^2 + \frac{[SO_2]}{(7.0\,x\,10^{-9})}(1.94\,\text{x}\,10^{-8}\,x\,GlobRad + 1.44\,\text{x}\,10^{-29}[O_3][Alkene])\right]^{1/2} \quad (12)$$

Table ~~1~~2: Coefficients used in the proxy equation in all four environments. Numbers in parenthesis represent the 25th and 75th percentiles of boot strapped data, respectively. See supplementary section 2 for more details.

~~Location~~	~~*GlobRad (W/m²)*~~	~~k_1(10⁻⁸ m² W⁻¹ s⁻¹)~~	~~k_2(·10⁻²⁹ cm⁶ s⁻¹)~~	~~k_3 (·10⁻⁹ cm³ s⁻¹)~~
~~Hyytiälä~~	~~>=0~~	~~1.21(1.15-1.24)~~	~~10.3(10.0-10.61)~~	~~5.98(5.58-5.99)~~
~~Agia Marina~~	~~>= 50~~	~~0.92(0.78-1.13)~~	~~N/A~~	~~2.32(1.47-3.63)~~
~~Budapest~~	~~>= 50~~	~~0.14(0.13-0.15)~~	~~N/A~~	~~7.90(7.90-7.91)~~
~~Beijing~~	~~>= 50~~	~~5.20(4.62 – 5.78)~~	~~1.45(1.09 – 1.88)~~	~~5.76(4.30 – 7.0)~~
~~Beijing~~	~~<50~~	~~1.35(1.09 – 1.64)~~	~~4.39(4.24 – 4.59)~~	~~7.0(6.99 – 7.0)~~

Location	*GlobRad (W/m²)*	k_1(10⁻⁸ m² W⁻¹ s⁻¹)	k_2(·10⁻²⁹ cm⁶ s⁻¹)	k_3 (·10⁻⁹ cm³ s⁻¹)
Hyytiälä	>0	0.85(0.60-1.21)	6.10(4.27-8.57)	4.26(2.98-5.99)
Agia Marina	>= 50	0.92(0.64-1.34)	N/A	2.21(1.27-3.79)
Budapest	>= 50	0.16(0.09-0.27)	N/A	9.80(9.79-9.81)
Beijing	> 0	1.94(1.12 – 3.50)	1.45(0.93 – 2.26)	7.0

Table ~~2~~3: Fraction of each source and sink term to the change in ~~otal~~ H_2SO_4 concentration. Median of boot ~~strapping~~ strap resampling results and their 25th and 75~~27~~th percentiles are shown.

	~~GlobRad (W/m²)~~	~~Source Terms~~		~~Sink Terms~~	
		~~k_1Glob[SO_2]~~	~~k_2[O_3][A][SO_2]~~	~~$-k_3[H_2SO_4]^2$~~	~~$-CS[H_2SO_4]$~~
~~Hyytiälä~~	~~>=0~~	~~0.31 (0.08-0.43)~~	~~0.18 (0.06-0.41)~~	~~0.16 (0.06-0.29)~~	~~0.34 (0.21-0.44)~~

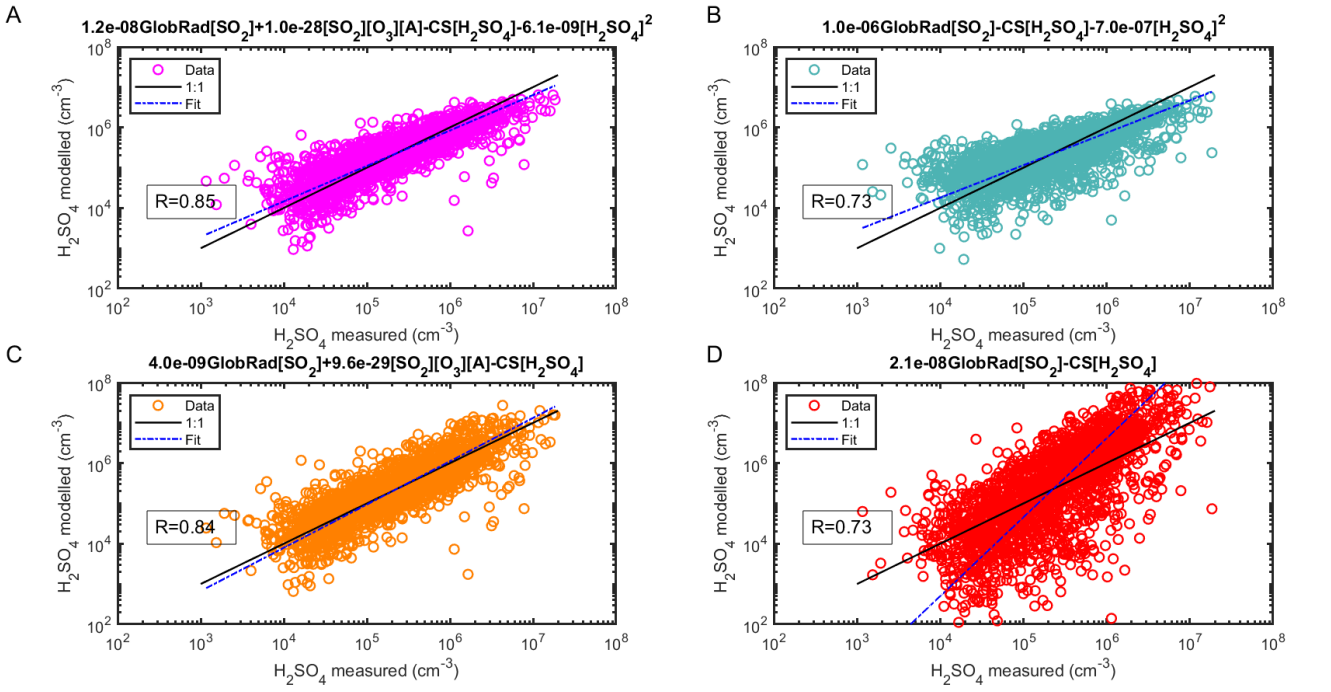
~~Agia Marina~~	~~>= 50~~	~~0.5 (0.48-0.52)~~	~~0~~	~~0.20 (0.15-0.32)~~	~~0.30 (0.18-0.33)~~
~~Budapest~~	~~>= 50~~	~~0.5 (0.48-0.51)~~	~~0~~	~~0.22 (0.15-0.29)~~	~~0.28 (0.21-0.35)~~
~~Beijing~~	~~>= 50~~	~~0.29 (0.24 – 0.35)~~	~~0.21 (0.15 – 0.26)~~	~~0.29 (0.18 – 0.34)~~	~~0.21 (0.14 – 0.30)~~
	~~<50~~	~~0.06 (0.02 – 0.13)~~	~~0.44 (0.36 – 0.48)~~	~~0.24 (0.11-0.35)~~	~~0.26 (0.15-0.39)~~

716

	GlobRad (W/m^2)	Source Terms		Sink Terms	
		$k_1 Glob[SO_2]$	$k_2[O_3][A][SO_2]$	$-k_3[H_2SO_4]^2$	$-CS[H_2SO_4]$
Hyytiälä	>0	0.34 (0.10-0.44)	0.16 (0.08-0.40)	0.16 (0.08-0.26)	0.34 (0.24-0.42)
Agia Marina	>= 50	0.5	0	0.24 (0.19-0.29)	0.26 (0.21-0.31)
Budapest	>= 50	0.5	0	0.26 (0.18-0.31)	0.24 (0.19-0.32)
Beijing	> 0	0.28 (2E-4– 0.41)	0.22 (0.09 – 0.50)	0.29 (0.19 – 0.39)	0.21 (0.11 – 0.31)

717

718



719

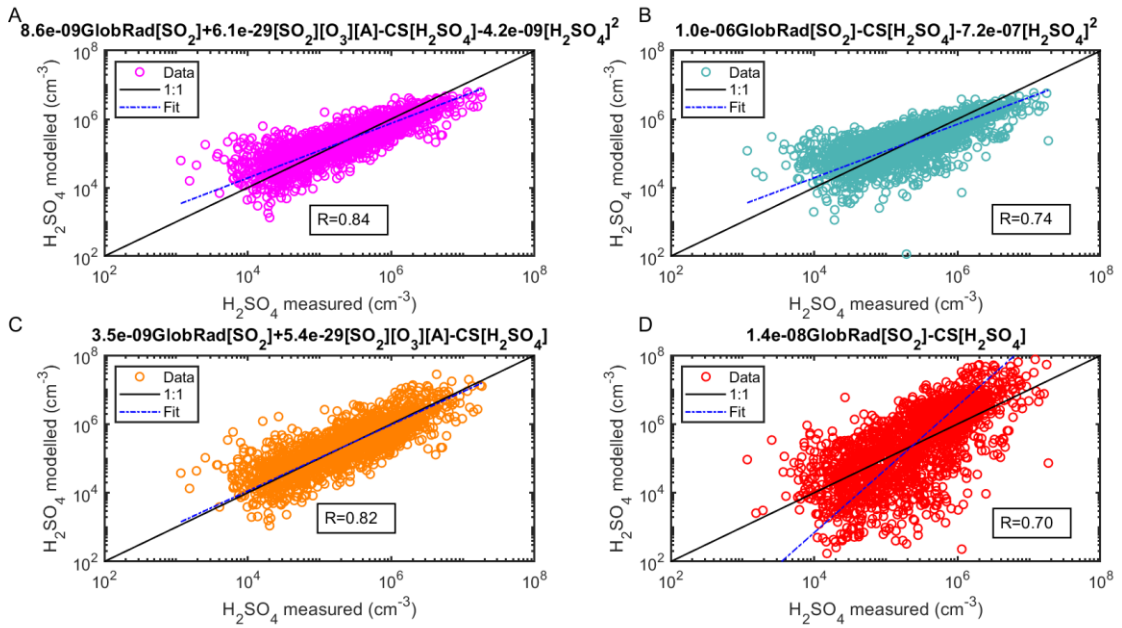
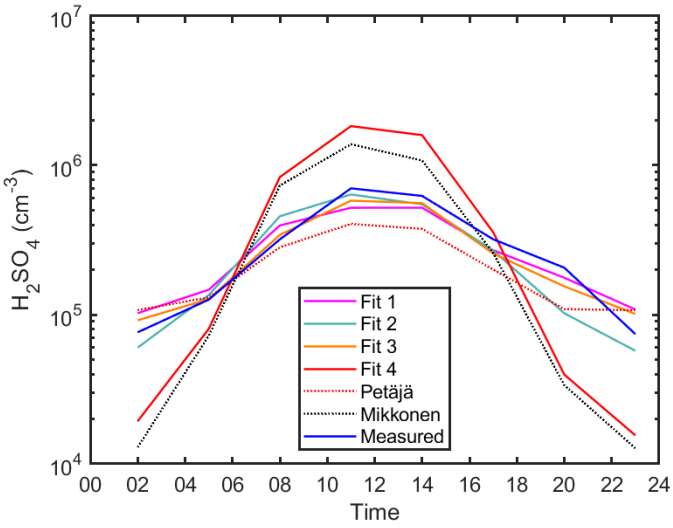


Figure 1: Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at SMEAR II station, Hyytiälä Finland. The observed concentrations from the training data set are measured 2016-2019 using CI-APi-ToF and are ~~1~~3-hour medians resulting in a total of ~~2089~~ 1860 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).



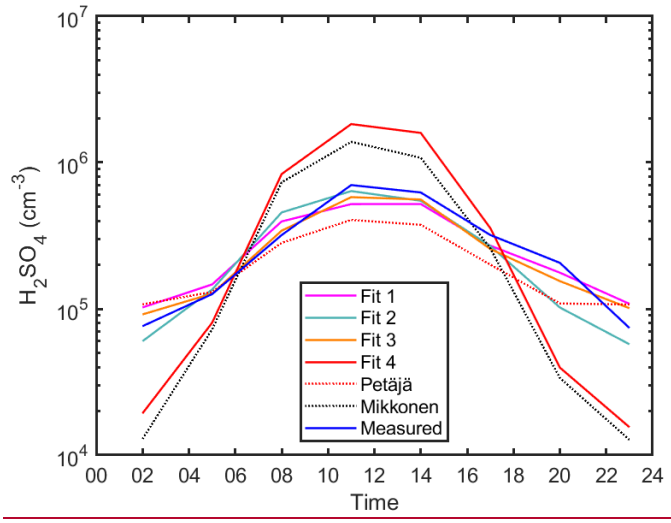


Figure 2: The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR II in Hyytiälä, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

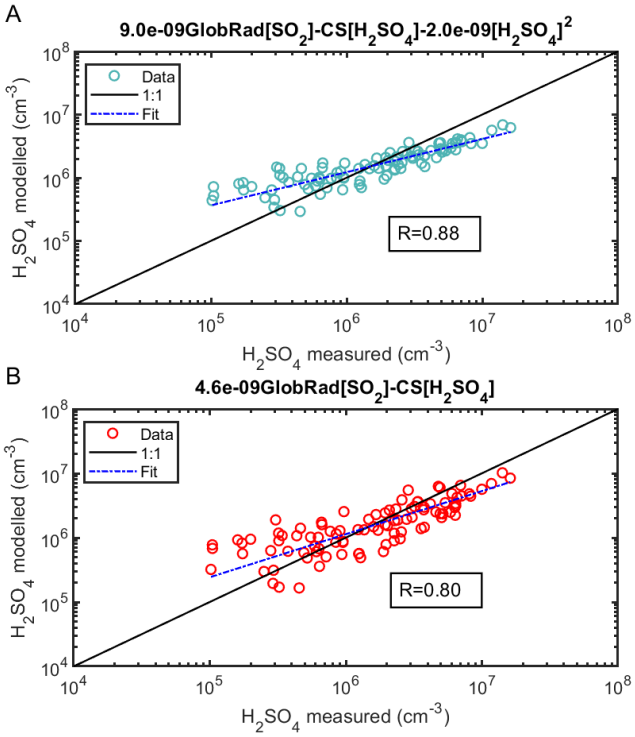


Figure 3: Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at Agia Marina, Cyprus, excluding the Alkene term. The observed numbers concentrations are measured during Feb- Mar 2018 using CI-APi-ToF and are hourly medians resulting in a total of 96 data points. Sulphuric acid proxy concentration as a function of measured sulphuric acid. In (A), the equation without the Stabilized Criegee Intermediates source (Equation 4) and in (B) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

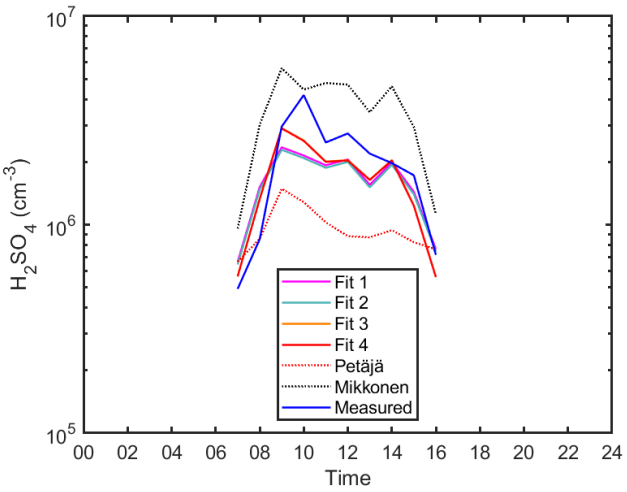


Figure 4 The diurnal variation of sulphuric acid proxies and observed concentrations in Agia Marina, Cyprus. Hourly median values are shown. Fits 2 and 4 corresponds to the Equations 4 and 6, respectively, See also Figure 3A and B, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

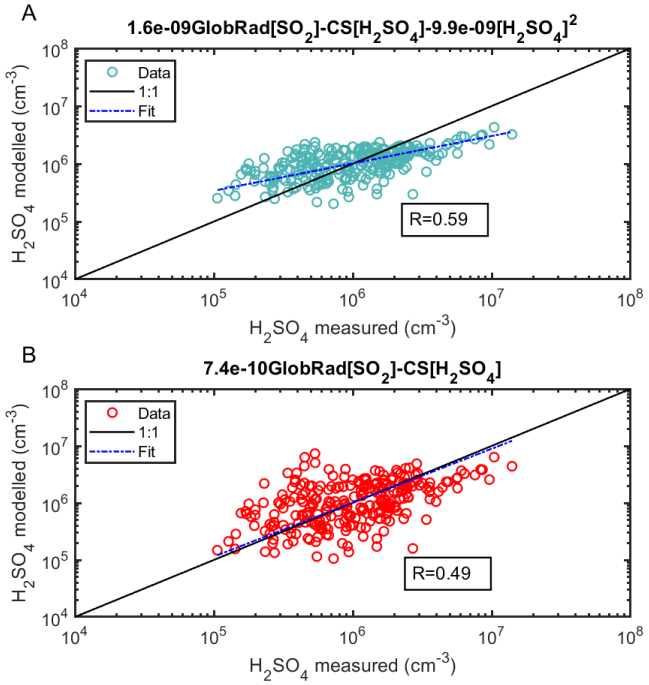


Figure 5 Sulphuric acid proxy as a function measured sulphuric acid at Budapest station, excluding the Alkene term. The observed numbers are measured during spring 2018 using CI-APi-ToF and are 1-hour medians coinciding with the measurement of trace gases and Global radiation every one hour resulting in a total of ~~262~~ 263 data points. In (A), the equation without the Stabilized Criegee Intermediates source (Equation 4) and in (B) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

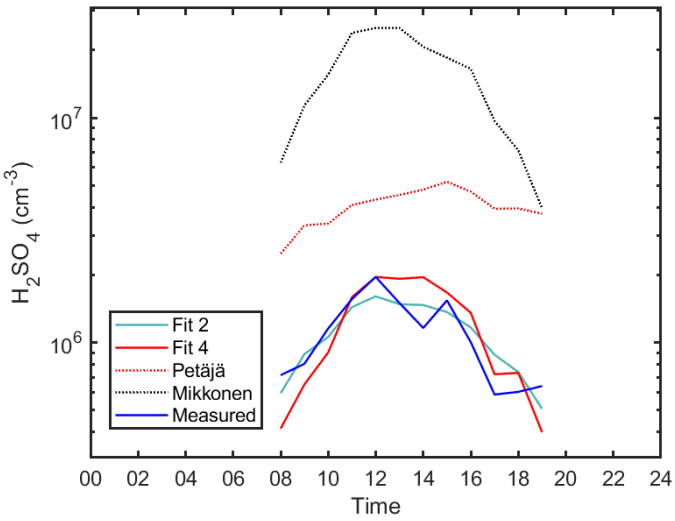
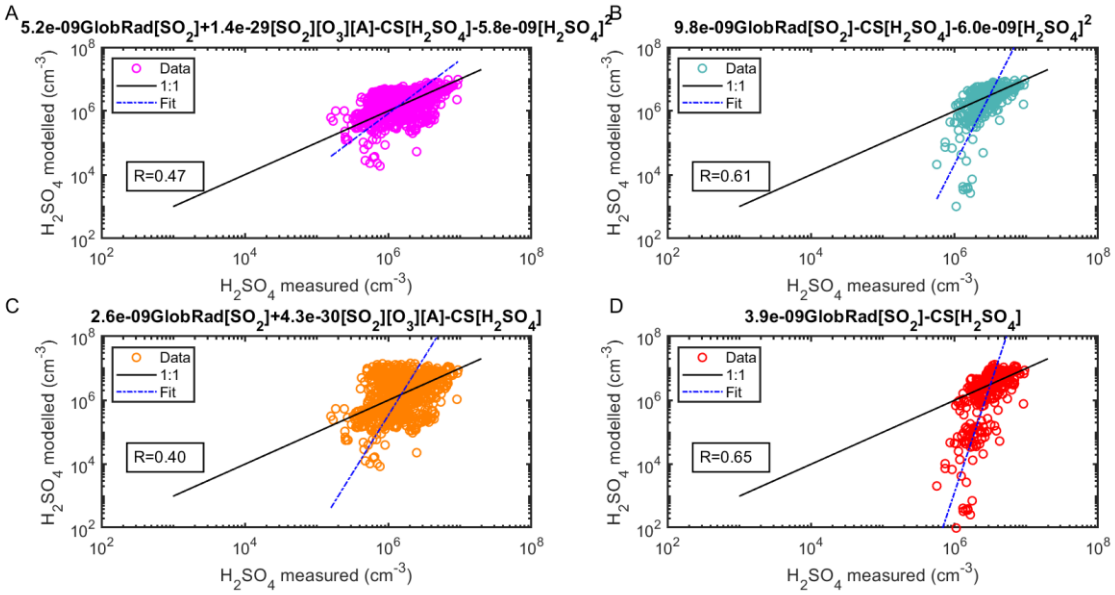


Figure 6 The diurnal variation of sulphuric acid proxies and measured concentrations in Budapest. Hourly median values are shown. Fits 2 and 4 corresponds to the Equations 4 and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).



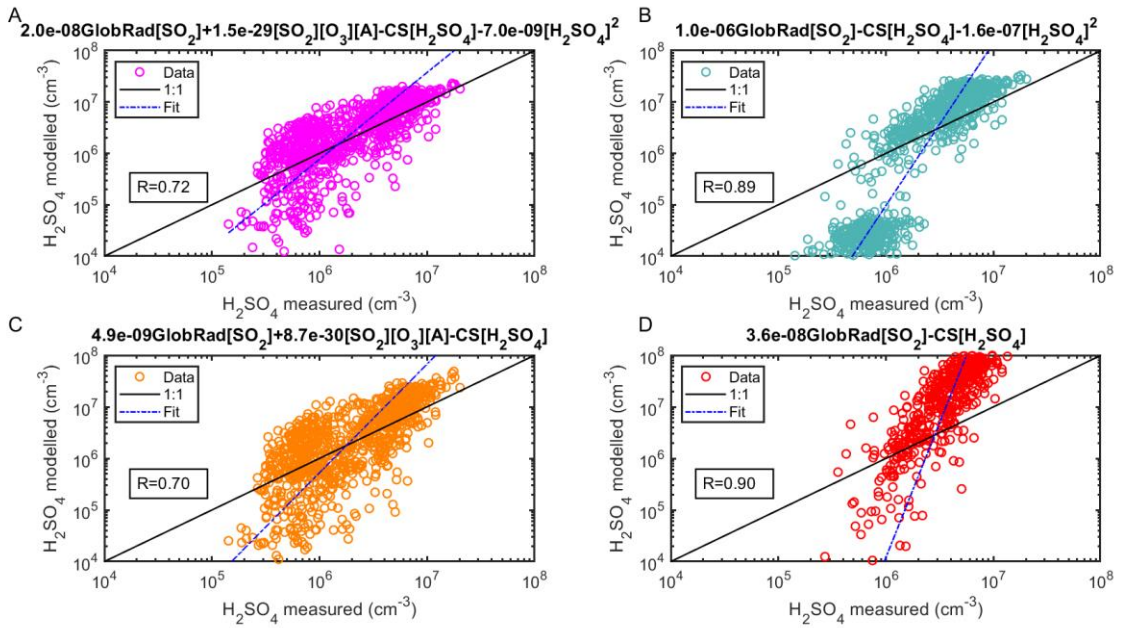


Figure 7 (A) Sulphuric acid proxy concentration ~~using Globrad~~ as a function of measured sulphuric acid. Observation at Beijing, China. The observed ~~numbers~~ concentrations of the training data set are measured ~~between 2018~~ in 2019 using CI-APi-ToF and are 1-hour medians resulting in a total of ~~875~~ 877 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). Coefficients shown on top of the subplots relate to the daytime values. The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

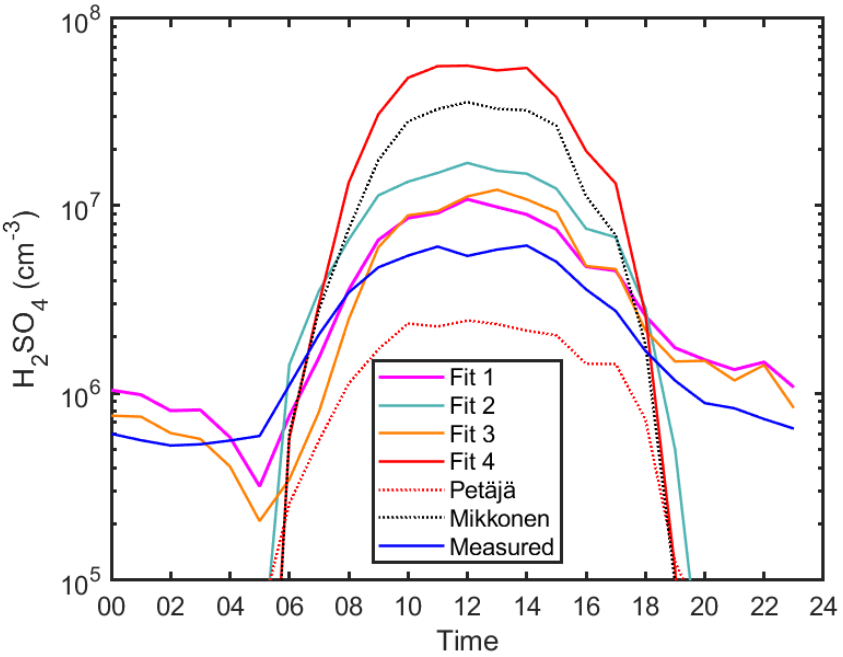


Figure 8 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at Beijing China, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

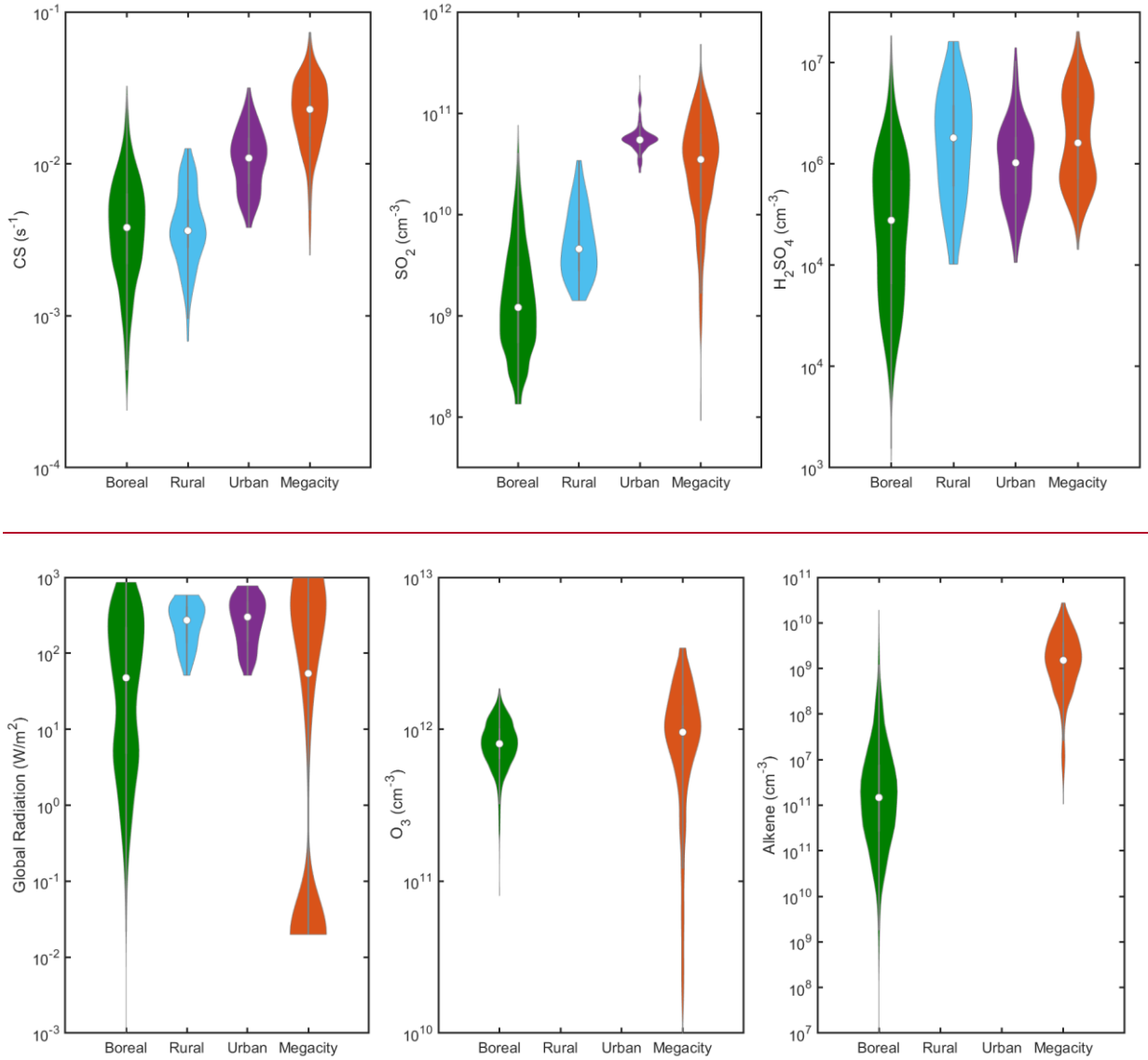


Figure 9 Characteristic predictor variables and H_2SO_4 concentrations in diffrerent environement.s O_3 and Alkenes data are available from the boreal forest (Hyytiälä) and megacity (Beijing) environments. This figure could be used in order to choose the equation and coefficients for calculating sulphuric acid proxy at a new location. The alkenes in the boreal environment are monoterpenes(e.g. alpha-pinene) and in the Megacity are anthropogenic volatile organic compounds (butylene, butadiene, isoprene, pentene and hexene). The concentrations are displayed as violin plots which are a combination of boxplot and a kernel distribution function on each side of the boxplots. The white circles define the median of the distribution and the edges on the inner grey boxes refer to the 25th and 75th percentiles respectively. Whole day data is shown for Hyytiala and Beijing, while daytime data (GlobRad > 50 W/m^2) for Agia Marina and Budapest. Daytime data (GlobRad > 50 W/m^2) is shown in Figure S15. The correlations between the different variables at each site are shown in Figures S2 – S6.

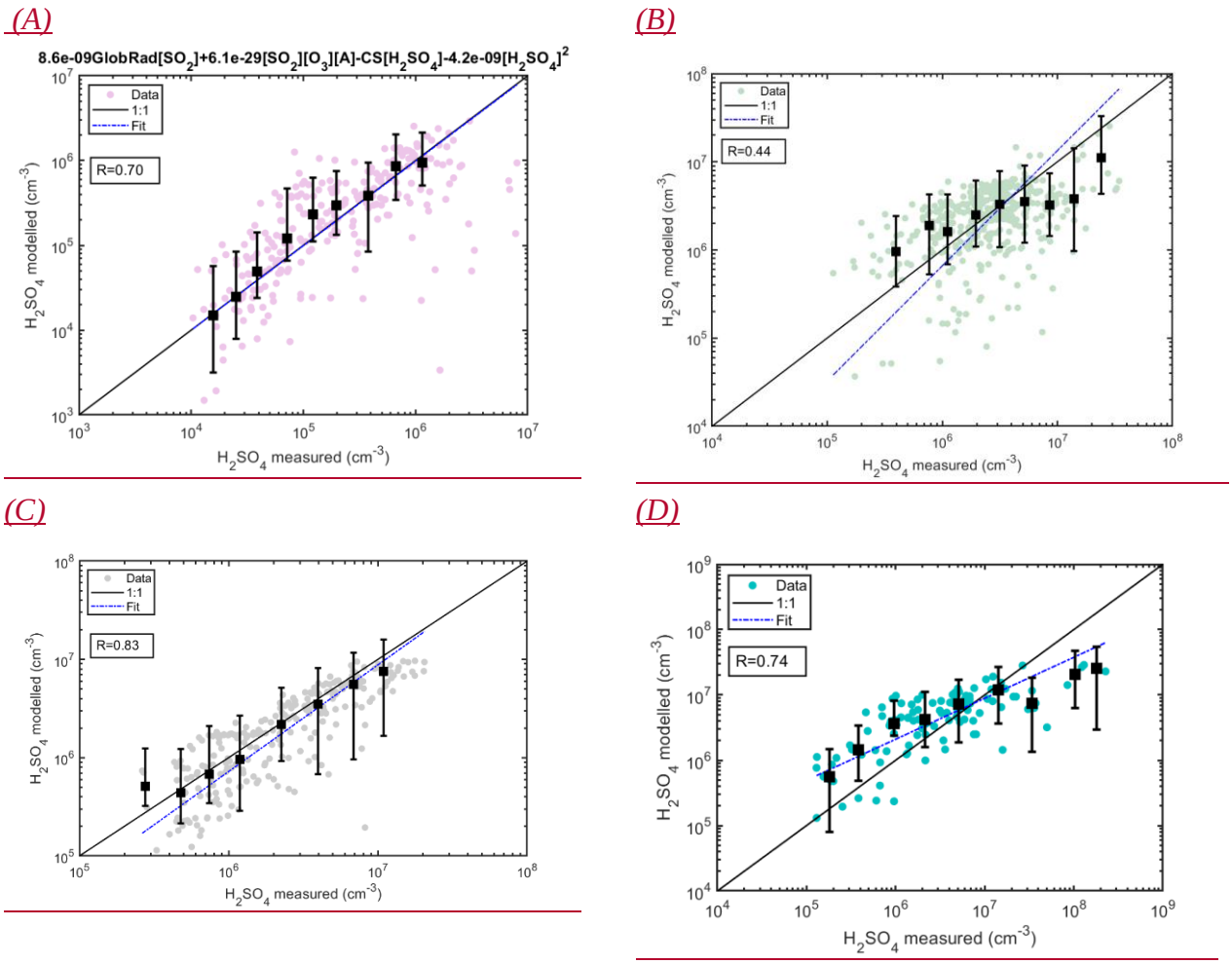


Figure 10 Sulphuric acid concentrations modelled as a function of measured sulphuric acid using testing data sets. The colored data points refer to the modelled (predicted) concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically-spaced measured sulphuric acid bins and their lower and upper whiskers correspond to 25th and 75th percentiles of the predicted concentrations. (A) Hyytiälä SMEAR II station: the concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. The modelled concentrations are derived using equation 9. (B) Helsinki SMEAR III station: the concentrations shown are 1-hour medians resulting in a total of 416 data points. The modelled concentrations are derived using equation 10. (C) Beijing: the concentrations shown are 1-hour medians resulting in a total of 268 data points. The modelled concentrations are derived using equation 12. (D) Kilpilahti: the concentrations shown are 1-hour medians resulting in 114 data points. The modelled concentrations are derived using equation 9.

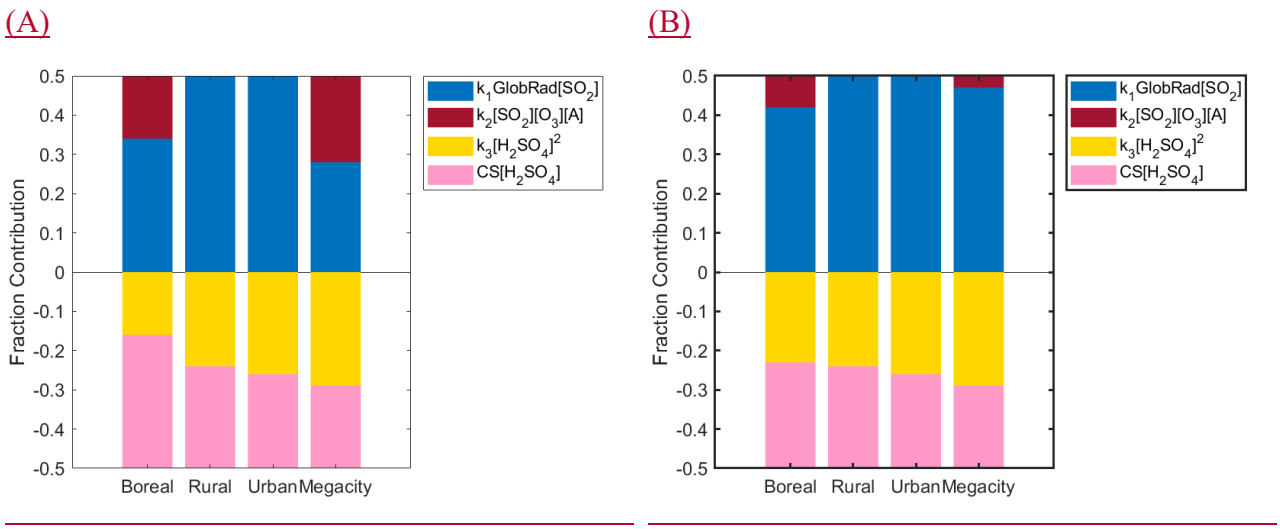
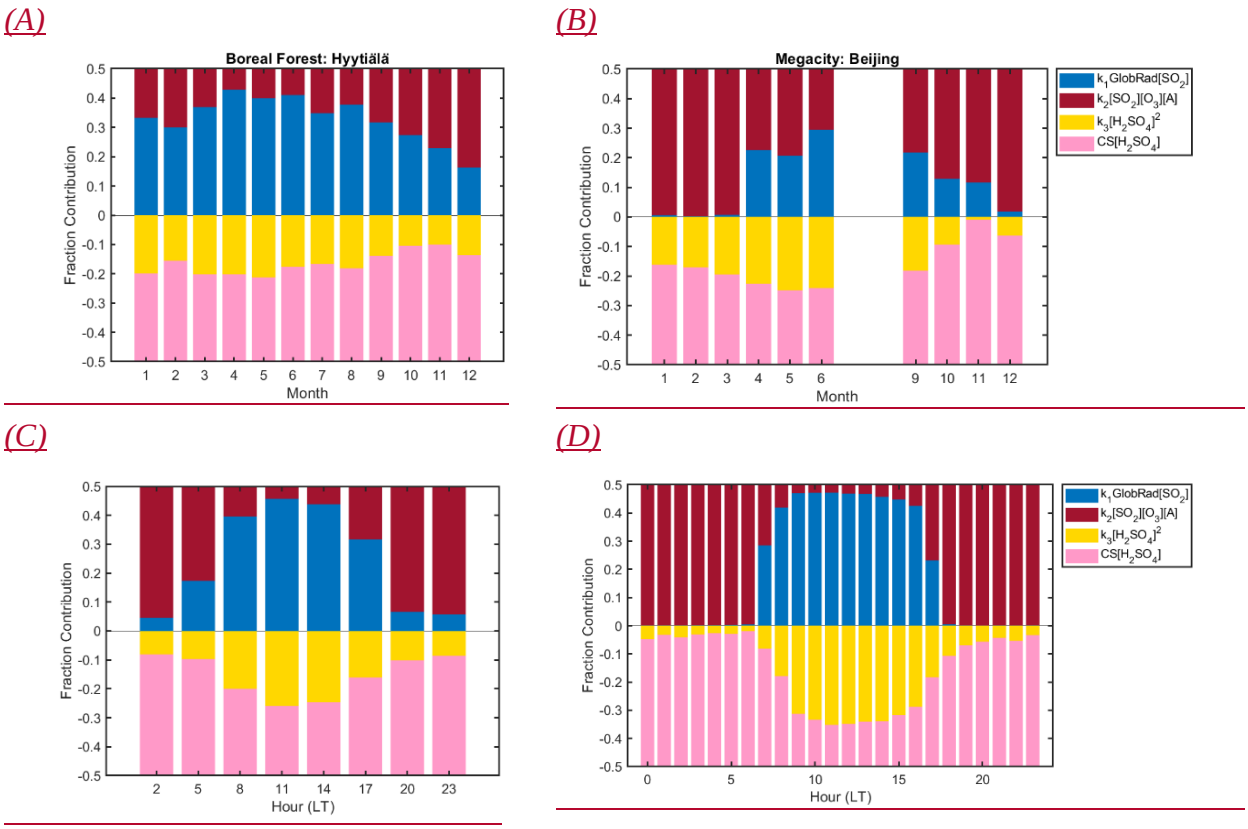
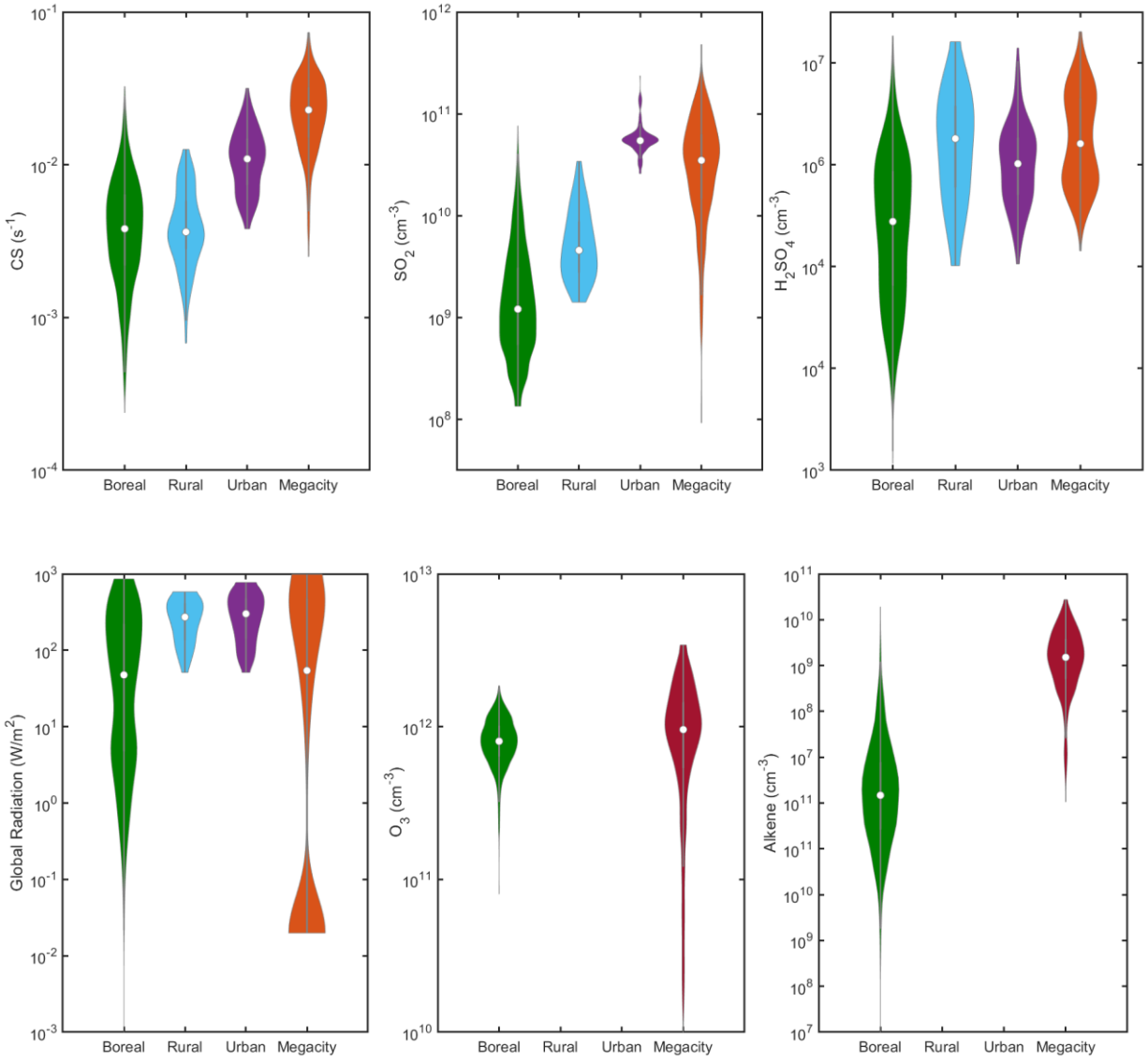


Figure 11 Fraction contribution of each source and sink term to the change in H_2SO_4 concentration. Figure ~~9~~ 11 is complementary to Table ~~2~~3. The boreal, rural, urban and megacity labels refer to Hyytiälä, Agia Marina, Budapest and Beijing sites, respectively. Note that the fraction of the alkene term contribution is not zero for the rural or urban sites, but is due to unavailable alkene data from these ~~Budapest and Cyprus~~ sites. In (A) we show all day medians for Hyytiälä and Beijing and in (B) we show daytime medians for all sites.



(A) Monthly variation of each source and sink term fraction contribution to the change in H_2SO_4 concentration in Hyytiälä within the training data set 2016-2019. (B) Monthly variation of each source and sink term to the change in H_2SO_4 concentration in Beijing within the training and testing data sets 2019, the data outside the training and testing data sets has missing measured sulphuric acid concentrations, so proxy concentrations were used in obtaining this figure. (C) Diurnal variation of each source and sink term to the change in H_2SO_4 concentration in Hyytiälä within the training data set. (D) Diurnal variation of each source and sink term to the change in H_2SO_4 concentration in Beijing within the training and testing data sets.



~~*Figure 10 Condensation Sink, SO_2 and H_2SO_4 concentrations in diffrerent environements and O_3 and Alkenes in the boreal forest (Hyytiälä) and megacity (Beijing) environments. This figure could be used in order to choose the coefficients for calculating the proxy. The alkenes in the boreal environment are monoterpenes(e.g. alpha pinene) and in the Megacity are anthropogenic volatile organic compounds (propylene, butylene, butadiene, isoprene, pentene and hexene). The concentrations are displayed as violin plots which are a combination of boxplot and a kernel distribution function on each side of the boxplots. The white circles define the median of the distribution and the edges on the inner grey boxes refer to the 25th and 75th percentiles respectively. Daytime data (GlobRad > 50 W/m^2) is shown in Figure S11-S12. The correlations between the different variables at each site are shown in Figures S3 – S7.*~~

References

Aalto, P., Hameri, K., Becker, E., Weber, R., Salm, J., Makela, J. M., Hoell, C., O'Dowd, C. D., Karlsson, H., Hansson, H. C., Vakeva, M., Koponen, I. K., Buzorius, G., and Kulmala, M.: Physical characterization of aerosol particles during nucleation events, Tellus B, 53, 344-358, DOI 10.1034/j.1600-0889.2001.530403.x, 2001.

Almeida, J., Schobesberger, S., Kurten, A., Ortega, I. K., Kupiainen-Maatta, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurten, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppa, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petaja, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipila, M., Stozhkov, Y., Stratmann, F., Tome, A., Trostl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P. L., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamaki, H., and Kirkby, J.: Molecular understanding of sulphuric acid-amine particle nucleation in the atmosphere, Nature, 502, 359-363, 10.1038/nature12663, 2013.

Baalbaki, R., Pikridas M., Jokinen T., Dada L., Ahonen L., Lehtipalo K., Petäjä T., Sciare J. Kulmala M.: Towards understanding the mechanisms of new particle formation in the Eastern Mediterranean, 2020, In Prep.

Berresheim, H., Elste, T., Tremmel, H. G., Allen, A. G., Hansson, H. C., Rosman, K., Dal Maso, M., Makela, J. M., Kulmala, M., and O'Dowd, C. D.: Gas-aerosol relationships of H2SO4, MSA, and OH: Observations in the coastal marine boundary layer at Mace Head, Ireland, J Geophys Res-Atmos, 107, Artn 8100 10.1029/2000jd000229, 2002.

Dada, L., Paasonen, P., Nieminen, T., Mazon, S. B., Kontkanen, J., Perakyla, O., Lehtipalo, K., Hussein, T., Petaja, T., Kerminen, V. M., Back, J., and Kulmala, M.: Long-term analysis of clear-sky new particle formation events and nonevents in Hyytiala, Atmos Chem Phys, 17, 6227-6241, 10.5194/acp-17-6227-2017, 2017.

Dada, L., Chellapermal, R., Buenrostro Mazon, S., Paasonen, P., Lampilahti, J., Manninen, H. E., Junninen, H., Petäjä, T., Kerminen, V. M., and Kulmala, M.: Refined classification and characterization of atmospheric new-particle formation events using air ions, Atmos. Chem. Phys., 18, 17883-17893, 10.5194/acp-18-17883-2018, 2018.

Deng, C., Fu, Y., Dada, L., Yan, C., Cai, R., Yang, D., Zhou, Y., Yin, R., Lu, Y., Li, X., Qiao, X., Fan, X., Nie, W., Kontkanen, J., Kangasluoma, J., Chu, B., Ding, A., Kerminen, V.-M., Paasonen, P., Worsnop, D. R., Bianchi, F., Liu, Y., Zheng, J., Wang, L., Kulmala, M., and Jiang, J.: Seasonal Characteristics of New Particle Formation and Growth in Urban Beijing, Environ Sci Technol, 10.1021/acs.est.0c00808, 2020.

Dunne, E. M., Gordon, H., Kurten, A., Almeida, J., Duplissy, J., Williamson, C., Ortega, I. K., Pringle, K. J., Adamov, A., Baltensperger, U., Barmet, P., Benduhn, F., Bianchi, F., Breitenlechner, M., Clarke, A., Curtius, J., Dommen, J., Donahue, N. M., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Jokinen, T.,

Kangasluoma, J., Kirkby, J., Kulmala, M., Kupc, A., Lawler, M. J., Lehtipalo, K., Makhmutov, V., Mann, G., Mathot, S., Merikanto, J., Miettinen, P., Nenes, A., Onnela, A., Rap, A., Reddington, C. L. S., Riccobono, F., Richards, N. A. D., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Simon, M., Sipilaa, M., Smith, J. N., Stozkhov, Y., Tome, A., Trostl, J., Wagner, P. E., Wimmer, D., Winkler, P. M., Worsnop, D. R., and Carslaw, K. S.: Global atmospheric particle formation from CERN CLOUD measurements, Science, 354, 1119-1124, 10.1126/science.aaf2649, 2016.

Efron, B., and Tibshirani, R. J.: An introduction to the bootstrap, CRC press, 1994.

Eisele, F. L., and Tanner, D. J.: Measurement of the Gas-Phase Concentration of H2so4 and Methane Sulfonic-Acid and Estimates of H2so4 Production and Loss in the Atmosphere, J Geophys Res-Atmos, 98, 9001-9010, Doi 10.1029/93jd00031, 1993.

Erupe, M. E., Viggiano, A. A., and Lee, S. H.: The effect of trimethylamine on atmospheric nucleation involving H2SO4, Atmos. Chem. Phys., 11, 4767-4775, 10.5194/acp-11-4767-2011, 2011.

Gao, W., Tan, G., Hong, Y., Li, M., Nian, H., Guo, C., Huang, Z., Fu, Z., Dong, J., Xu, X., Cheng, P., and Zhou, Z.: Development of portable single photon ionization time-of-flight mass spectrometer combined with membrane inlet, International Journal of Mass Spectrometry, 334, 8-12, https://doi.org/10.1016/j.ijms.2012.09.003, 2013.

Gordon, H., Kirkby, J., Baltensperger, U., Bianchi, F., Breitenlechner, M., Curtius, J., Dias, A., Dommen, J., Donahue, N. M., Dunne, E. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Frege, C., Fuchs, C., Hansel, A., Hoyle, C. R., Kulmala, M., Kurten, A., Lehtipalo, K., Makhmutov, V., Molteni, U., Rissanen, M. P., Stozkhov, Y., Trostl, J., Tsagkogeorgas, G., Wagner, R., Williamson, C., Wimmer, D., Winkler, P. M., Yan, C., and Carslaw, K. S.: Causes and importance of new particle formation in the present-day and preindustrial atmospheres, J Geophys Res-Atmos, 122, 8739-8760, 10.1002/2017jd026844, 2017.

Guo, S., Hu, M., Zamora, M. L., Peng, J., Shang, D., Zheng, J., Du, Z., Wu, Z., Shao, M., Zeng, L., Molina, M. J., and Zhang, R.: Elucidating severe urban haze formation in China, Proceedings of the National Academy of Sciences, 111, 17373-17378, 10.1073/pnas.1419604111 2014.

Hakola, H., Hellen, H., Hemmila, M., Rinne, J., and Kulmala, M.: In situ measurements of volatile organic compounds in a boreal forest, Atmos Chem Phys, 12, 11665-11678, 10.5194/acp-12-11665-2012, 2012.

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

Hellén, H., Praplan, A. P., Tykkä, T., Ylivinkka, I., Vakkari, V., Bäck, J., Petäjä, T., Kulmala, M., and Hakola, H.: Long-term measurements of volatile organic compounds highlight the importance of sesquiterpenes for the atmospheric chemistry of a boreal forest, Atmospheric Chemistry Physics
18, 13839-13863, 2018.

Hussein, T., Martikainen, J., Junninen, H., Sogacheva, L., Wagner, R., Dal Maso, M., Riipinen, I., Aalto, P. P., and Kulmala, M.: Observation of regional new particle formation in the urban atmosphere, Tellus B, 60, 509-521, 2008.

Jen, C. N., McMurry, P. H., and Hanson, D. R.: Stabilization of sulfuric acid dimers by ammonia, methylamine, dimethylamine, and trimethylamine, Journal of Geophysical Research: Atmospheres, 119, 7502-7514, 2014.

Jokinen, T., Sipila, M., Junninen, H., Ehn, M., Lonn, G., Hakala, J., Petaja, T., Mauldin, R. L., Kulmala, M., and Worsnop, D. R.: Atmospheric sulphuric acid and neutral cluster measurements using CI-APi-TOF, Atmos Chem Phys, 12, 4117-4125, 10.5194/acp-12-4117-2012, 2012.

Junninen, H., Ehn, M., Petaja, T., Luosujarvi, 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, Atmos Meas Tech, 3, 1039-1053, 10.5194/amt-3-1039-2010, 2010.

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

Kerminen, V.-M., Chen, X., Vakkari, V., Petäjä, T., Kulmala, M., and Bianchi, F.: Atmospheric new particle formation and growth: review of field observations, Environ. Res. Lett., 13, 103003, 10.1088/1748-9326/aadf3c, 2018.

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

Kulmala, M., Petaja, T., Nieminen, T., Sipila, 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, 10.1038/nprot.2012.091, 2012.

Kulmala, M., Kontkanen, J., Junninen, H., Lehtipalo, K., Manninen, H. E., Nieminen, T., Petaja, T., Sipila, M., Schobesberger, S., Rantala, P., Franchin, A., Jokinen, T., Jarvinen, E., Aijala, M., Kangasluoma, J., Hakala, J., Aalto, P. P., Paasonen, P., Mikkila, J., Vanhanen, J., Aalto, J., Hakola, H., Makkonen, U., Ruuskanen, T., Mauldin, R. L., Duplissy, J., Vehkamaki, H., Back, J., Kortelainen, A., Riipinen, I., Kurten, 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, 10.1126/science.1227385, 2013.

Kulmala, M., Kerminen, V. M., Petaja, T., Ding, A. J., and Wang, L.: Atmospheric gas-to-particle conversion: why NPF events are observed in megacities?, Faraday Discuss, 200, 271-288, 10.1039/c6fd00257a, 2017.

Kurten, A., Rondo, L., Ehrhart, S., and Curtius, J.: Calibration of a chemical ionization mass spectrometer for the measurement of gaseous sulfuric acid, Phys Chem A, 116, 6375-6386, 10.1021/jp212123n, 2012.

Kürten, A., Williamson, C., Almeida, J., Kirkby, J., and Curtius, J.: On the derivation of particle nucleation rates from experimental formation rates, Atmos. Chem. Phys., 15, 4063-4075, 10.5194/acp-15-4063-2015, 2015.

Laakso, L., Petaja, T., Lehtinen, K. E. J., Kulmala, M., Paatero, J., Horrak, U., Tammet, H., and Joutsensaari, J.: Ion production rate in a boreal forest based on ion, particle and radiation measurements, Atmos Chem Phys, 4, 1933-1943, DOI 10.5194/acp-4-1933-2004, 2004.

Laakso, L., Koponen, I. K., Monkkonen, P., Kulmala, M., Kerminen, V. M., Wehner, B., Wiedensohler, A., Wu, Z. J., and Hu, M.: Aerosol particles in the developing world; A comparison between New Delhi in India and Beijing in China, Water Air Soil Poll, 173, 5-20, 10.1007/s11270-005-9018-5, 2006.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E.: Convergence Properties of the Nelder--Mead Simplex Method in Low Dimensions, SIAM Journal on Optimization, 9, 112-147, 10.1137/s1052623496303470, 1998.

Lehtipalo, K., Yan, C., Dada, L., Bianchi, F., Xiao, M., Wagner, R., Stolzenburg, D., Ahonen, L. R., Amorim, A., Baccarini, A., Bauer, P. S., Baumgartner, B., Bergen, A., Bernhammer, A.-K., Breitenlechner, M., Brilke, S., Buchholz, A., Mazon, S. B., Chen, D., Chen, X., Dias, A., Dommen, J., Draper, D. C., Duplissy, J., Ehn, M., Finkenzeller, H., Fischer, L., Frege, C., Fuchs, C., Garmash, O., Gordon, H., Hakala, J., He, X., Heikkinen, L., Heinritzi, M., Helm, J. C., Hofbauer, V., Hoyle, C. R., Jokinen, T., Kangasluoma, J., Kerminen, V.-M., Kim, C., Kirkby, J., Kontkanen, J., Kürten, A., Lawler, M. J., Mai, H., Mathot, S., Mauldin, R. L., Molteni, U., Nichman, L., Nie, W., Nieminen, T., Ojdanic, A., Onnela, A., Passananti, M., Petäjä, T., Piel, F., Pospisilova, V., Quéléver, L. L. J., Rissanen, M. P., Rose, C., Sarnela, N., Schallhart, S., Schuchmann, S., Sengupta, K., Simon, M., Sipilä, M., Tauber, C., Tomé, A., Tröstl, J., Väisänen, O., Vogel, A. L., Volkamer, R., Wagner, A. C., Wang, M., Weitz, L., Wimmer, D., Ye, P., Ylisirniö, A., Zha, Q., Carslaw, K. S., Curtius, J., Donahue, N. M., Flagan, R. C., Hansel, A., Riipinen, I., Virtanen, A., Winkler, P. M., Baltensperger, U., Kulmala, M., and Worsnop, D. R.: Multicomponent new particle formation from sulfuric acid, ammonia, and biogenic vapors, Science Advances, 4, eaau5363, 10.1126/sciadv.aau5363 2018.

Liu, J., Jiang, J., Zhang, Q., Deng, J., and Hao, J.: A spectrometer for measuring particle size distributions in the range of 3 nm to 10 μm, J Frontiers of Environmental Science and Engineering, 10, 63-72, 10.1007/s11783-014-0754-x, 2016.

Lu, Y., Yan, C., Fu, Y., Chen, Y., Liu, Y., Yang, G., Wang, Y., Bianchi, F., Chu, B., Zhou, Y., Yin, R., Baalbaki, R., Garmash, O., Deng, C., Wang, W., Liu, Y., Petäjä, T., Kerminen, V. M., Jiang, J., Kulmala, M., and Wang, L.: A proxy for atmospheric daytime gaseous sulfuric acid concentration in urban Beijing, Atmos. Chem. Phys., 19, 1971-1983, 10.5194/acp-19-1971-2019, 2019.

Ma, F., Xie, H.-B., Elm, J., Shen, J., Chen, J., and Vehkamäki, H.: Piperazine Enhancing Sulfuric Acid-Based New Particle Formation: Implications for the Atmospheric Fate of Piperazine, Environ Sci Technol, 53, 8785-8795, 10.1021/acs.est.9b02117, 2019.

Mauldin, R. L., Berndt, T., Sipila, M., Paasonen, P., Petaja, T., Kim, S., Kurten, T., Stratmann, F., Kerminen, V. M., and Kulmala, M.: A new atmospherically relevant oxidant of sulphur dioxide, Nature, 488, 193-196, 10.1038/nature11278, 2012.

McElreath, R.: Statistical rethinking: A Bayesian course with examples in R and Stan, Chapman and Hall/CRC, 2018.

Merikanto, J., Spracklen, D. V., Mann, G. W., Pickering, S. J., and Carslaw, K. S.: Impact of nucleation on global CCN, Atmos Chem Phys, 9, 8601-8616, 10.5194/acp-9-8601-2009, 2009.

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petaja, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M., and Laaksonen, A.: A statistical proxy for sulphuric acid concentration, Atmos Chem Phys, 11, 11319-11334, 10.5194/acp-11-11319-2011, 2011.

Mikkonen, S., Németh, Z., Varga, V., Weidinger, T., Leinonen, V., Yli-Juuti, T., and Salma, I.: Decennial time trends and diurnal patterns of particle number concentrations in a Central European city between 2008 and 2018, Atmos. Chem. Phys. Discuss., 2020, 1-27, 10.5194/acp-2020-305, 2020.

Monkkonen, P., Koponen, I. K., Lehtinen, K. E. J., Uma, R., Srinivasan, D., Hameri, K., and Kulmala, M.: Death of nucleation and Aitken mode particles: observations at extreme atmospheric conditions and their theoretical explanation, J Aerosol Sci, 35, 781-787, 10.1016/j.jaerosci.2003.12.004, 2004.

Monkkonen, P., Koponen, I. K., Lehtinen, K. E. J., Hameri, K., Uma, R., and Kulmala, M.: Measurements in a highly polluted Asian mega city: observations of aerosol number size distribution, modal parameters and nucleation events, Atmos Chem Phys, 5, 57-66, 2005.

Nieminen, T., Kerminen, V. M., Petäjä, T., Aalto, P. P., Arshinov, M., Asmi, E., Baltensperger, U., Beddows, D. C. S., Beukes, J. P., Collins, D., Ding, A., Harrison, R. M., Henzing, B., Hooda, R., Hu, M., Hõrrak, U., Kivekäs, N., Komsaare, K., Krejci, R., Kristensson, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Lihavainen, H., Mihalopoulos, N., Németh, Z., Nie, W., O'Dowd, C., Salma, I., Sellegri, K., Svenningsson, B., Swietlicki, E., Tunved, P., Ulevicius, V., Vakkari, V., Vana, M., Wiedensohler, A., Wu, Z., Virtanen, A., and Kulmala, M.: Global analysis of continental boundary layer new particle formation based on long-term measurements, Atmos. Chem. Phys., 18, 14737-14756, 10.5194/acp-18-14737-2018, 2018.

Novelli, A., Hens, K., Ernest, C. T., Martinez, M., Nolscher, A. C., Sinha, V., Paasonen, P., Petaja, T., Sipila, M., Elste, T., Plass-Dulmer, C., Phillips, G. J., Kubistin, D., Williams, J., Vereecken, L., Lelieveld, J., and Harder, H.: Estimating the atmospheric concentration of Criegee intermediates and their possible interference in a FAGE-LIF instrument, Atmos Chem Phys, 17, 7807-7826, 10.5194/acp-17-7807-2017, 2017.

Petäjä, T., Mauldin, 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, DOI 10.5194/acp-9-7435-2009, 2009.

Pikridas, M., Vrekoussis, M., Sciare, J., Kleanthous, S., Vasiliadou, E., Kizas, C., Savvides, C., and Mihalopoulos, N.: Spatial and temporal (short and long-term) variability of submicron, fine and sub-10 μm particulate matter (PM1, PM2.5, PM10) in Cyprus, Atmos Environ, 191, 79-93, https://doi.org/10.1016/j.atmosenv.2018.07.048, 2018.

Rinne, J., Ruuskanen, T. M., Reissell, A., Taipale, R., Hakola, H., and Kulmala, M.: On-line PTR-MS measurements of atmospheric concentrations of volatile organic compounds in a European boreal forest ecosystem, Boreal Environ Res, 10, 425-436, 2005.

Rohrer, F., and Berresheim, H.: Strong correlation between levels of tropospheric hydroxyl radicals and solar ultraviolet radiation, Nature, 442, 184-187, 10.1038/nature04924, 2006.

Salma, I., Németh, Z., Kerminen, V.-M., Aalto, P., Nieminen, T., Weidinger, T., Molnár, Á., Imre, K., and Kulmala, M.: Regional effect on urban atmospheric nucleation, Atmos Chem Phys, 16, 8715-8728, 2016a.

Salma, I., Németh, Z., Weidinger, T., Kovács, B., and Kristóf, G.: Measurement, growth types and shrinkage of newly formed aerosol particles at an urban research platform, Atmos. Chem. Phys., 16, 7837-7851, 10.5194/acp-16-7837-2016, 2016b.

Salma, I., and Németh, Z.: Dynamic and timing properties of new aerosol particle formation and consecutive growth events, Atmos. Chem. Phys., 19, 5835-5852, 10.5194/acp-19-5835-2019, 2019.

Sarnela, N., Jokinen, T., Nieminen, T., Lehtipalo, K., Junninen, H., Kangasluoma, J., Hakala, J., Taipale, R., Schobesberger, S., Sipila, M., Larnimaa, K., Westerholm, H., Heijari, J., Kerminen, V. M., Petaja, T., and Kulmala, M.: Sulphuric acid and aerosol particle production in the vicinity of an oil refinery, Atmos Environ, 119, 156-166, 10.1016/j.atmosenv.2015.08.033, 2015.

Sihto, S. L., Kulmala, M., Kerminen, V. M., Dal Maso, M., Petaja, T., Riipinen, I., Korhonen, H., Arnold, F., Janson, R., Boy, M., Laaksonen, A., and Lehtinen, K. E. J.: Atmospheric sulphuric acid and aerosol formation: implications from atmospheric measurements for nucleation and early growth mechanisms, Atmos Chem Phys, 6, 4079-4091, DOI 10.5194/acp-6-4079-2006, 2006.

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, 10.1126/science.1180315 2010.

Sipilä, M., Jokinen, T., Berndt, T., Richters, S., Makkonen, R., Donahue, N. M., Mauldin, R. L., Kurten, T., Paasonen, P., Sarnela, N., Ehn, M., Junninen, H., Rissanen, M. P., Thornton, J., Stratmann, F., Herrmann, H., Worsnop, D. R., Kulmala, M., Kerminen, V. M., and Petäjä, T.: Reactivity of stabilized Criegee intermediates (sCIs) from isoprene and monoterpene ozonolysis toward SO2 and organic acids, Atmos Chem Phys, 14, 12143-12153, 10.5194/acp-14-12143-2014, 2014.

Spracklen, D. V., Carslaw, K. S., Kulmala, M., Kerminen, V. M., Sihto, S. L., Riipinen, I., Merikanto, J., Mann, G. W., Chipperfield, M. P., and Wiedensohler, A.: Contribution of particle formation to global cloud condensation nuclei concentrations, Geophy. Res. Lett., 35, 2008.

Spracklen, D. V., Carslaw, K. S., Merikanto, J., Mann, G. W., Reddington, C. L., Pickering, S., Ogren, J. A., Andrews, E., Baltensperger, U., Weingartner, E., Boy, M., Kulmala, M., Laakso, L., Lihavainen, H., Kivekas, N., Komppula, M., Mihalopoulos, N., Kouvarakis, G., Jennings, S. G., O'Dowd, C., Birmili, W., Wiedensohler, A., Weller, R., Gras, J., Laj, P., Sellegri, K., Bonn, B., Krejci, R., Laaksonen, A., Hamed, A., Minikin, A., Harrison, R. M.,

Talbot, R., and Sun, J.: Explaining global surface aerosol number concentrations in terms of primary emissions and particle formation, Atmos Chem Phys, 10, 4775-4793, 10.5194/acp-10-4775-2010, 2010.

Taipale, R., Ruuskanen, T. M., Rinne, J., Kajos, M. K., Hakola, H., Pohja, T., and Kulmala, M.: Technical Note: Quantitative long-term measurements of VOC concentrations by PTR-MS - measurement, calibration, and volume mixing ratio calculation methods, Atmos Chem Phys, 8, 6681-6698, DOI 10.5194/acp-8-6681-2008, 2008.

Weber, R. J., Marti, J. J., McMurry, P. H., Eisele, F. L., Tanner, D. J., and Jefferson, A.: MEASURED ATMOSPHERIC NEW PARTICLE FORMATION RATES: IMPLICATIONS FOR NUCLEATION MECHANISMS, Chemical Engineering Communications, 151, 53-64, 10.1080/00986449608936541, 1996.

Yan, C., Dada, L., Rose, C., Jokinen, T., Nie, W., Schobesberger, S., Junninen, H., Lehtipalo, K., Sarnela, N., Makkonen, U., Garmash, O., Wang, Y., Zha, Q., Paasonen, P., Bianchi, F., Sipilä, M., Ehn, M., Petäjä, T., Kerminen, V. M., Worsnop, D. R., and Kulmala, M.: The role of H2SO4-NH3 anion clusters in ion-induced aerosol nucleation mechanisms in the boreal forest, Atmos. Chem. Phys., 18, 13231-13243, 10.5194/acp-18-13231-2018, 2018.

Yang, D., Zhang, S., Niu, T., Wang, Y., Xu, H., Zhang, K. M., and Wu, Y.: High-resolution mapping of vehicle emissions of atmospheric pollutants based on large-scale, real-world traffic datasets, Atmos. Chem. Phys., 19, 8831-8843, 10.5194/acp-19-8831-2019, 2019.

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, 10.1126/science.aao4839 2018.

Yao, L., Fan, X., Yan, C., Kurtén, T., Daellenbach, K. R., Wang, Y., Guo, Y., Li, C., Dada, L., Cai, J., Jun, T. Y., Zha, Q., Du, W., Yu, M., Zheng, F., Zhou, Y., Chan, T., Shen, J., Kujansuu, J. T., Kangasluoma, J., Jiang, J., Li, H., Wang, L., Worsnop, D. R., He, H., Petäjä, T., Kerminen, V.-M., Liu, Y., Chu, B., Kulmala, M., and Bianchi, F.: Enhanced atmospheric gaseous sulfuric acid formation from non-photochemical processes in urban Beijing, China, 2020, In Rev.

Zhang, R., Khalizov, A., Wang, L., Hu, M., and Xu, W.: Nucleation and growth of nanoparticles in the atmosphere, Chem. Rev., 112, 1957-2011, 2011.

Zhang, W., Tong, S., Ge, M., An, J., Shi, Z., Hou, S., Xia, K., Qu, Y., Zhang, H., Chu, B., Sun, Y., and He, H.: Variations and sources of nitrous acid (HONO) during a severe pollution episode in Beijing in winter 2016, Sci Total Environ, 648, 253-262, https://doi.org/10.1016/j.scitotenv.2018.08.133, 2019.

Zhou, Y., Dada, L., Liu, Y., Fu, Y., Kangasluoma, J., Chan, T., Yan, C., Chu, B., Daellenbach, K. R., Bianchi, F., Kokkonen, T. V., Liu, Y., Kujansuu, J., Kerminen, V. M., Petäjä, T., Wang, L., Jiang, J., and Kulmala, M.: Variation of size-segregated particle number concentrations in wintertime Beijing, Atmos. Chem. Phys., 20, 1201-1216, 10.5194/acp-20-1201-2020, 2020.

Supplementary Material for

Sources and sinks driving sulphuric acid concentrations in contrasting environments: implications on proxy calculations

by Lubna Dada *et al.*

1. Reaction rate constant from Mikkonen et al. 2011

Derivation of the temperature dependent reaction rate constant (k) used in calculating the Mikkonen proxy from our data sets:

$$k\ (cm^3\ molec^{-1}\ s^{-1}\ ) = \frac{A\ .k_3}{(A+k_3)}\ x \exp\left\{ k_5 \left[1 + \log_{10}\left(\frac{A}{k_3}\right)^2\right]^{-1} \right\} \quad (S1)$$

$$A = k_1 . [M] . \left(\frac{300}{k}\right)^{k_2} \quad (S2)$$

$$[M] = 0.101\ .\ (1.381\ x\ 10^{-23}\ T)^{-1} \quad (S3)$$

M is the density of the air in molec cm^{-3}, $k_1 = 4\times10^{-31}$, $k_2 = 3.3$, $k_3 = 2\times10^{-12}$ and $k_5 = -0.8$.
k given in Equation (S1) is scaled by multiplying it with 10^{12} as described in more detail in Mikkonen et al. (2011).

2. Bootstrap resampling and sensitivity analyses

When deriving the proxy equation for each site, 10 000 bootstrap resamples were drawn for each data set independently. Bootstrap resampling without disturbance generates extended data from the original data by randomly replacing an existing data point with another one from the same data set, resulting in different combinations of variables from the original data set. We accounted for the systematic uncertainty in H_2SO_4 and predictor variables arising e.g. from calibration uncertainties. For every bootstrap fit, we assumed both H_2SO_4 and all predictor variables to be affected by independent systematic errors between the upper and lower bound of their independent uncertainty ranges. Since the uncertainty related to the measurement accuracy was much larger than the precision of the measurement, we only accounted for the uncertainty arising from accuracy. In practice, we scaled the entire time series of each variable by a random set of numbers drawn from a uniform distribution of possible measurement biases.
Accordingly, a factor of 2 uncertainty was introduced in the sulphuric acid concentration, a 20% uncertainty in the condensation sink measurement, and a 10% in each trace gas concentration and global radiation. In the case of sulphuric acid concentrations, which have a factor of 2 uncertainty, the actual concentration of sulphuric acid at a certain point in time could be anywhere between a factor of 2 lower and a factor of 2 higher. Therefore, for each sulphuric acid measurement, we generated 10 000 concentrations by multiplying the original measured concentration by a uniform random array between the lower and upper bounds, which are 0.5 and 2 in the case of sulphuric acid. The same resampling method ws applied for each other predictor variable independently, and the 10 000 possible combinations of the disturbed data sets were used to generate the fit and to derive the sulphuric acid proxy equation per site. A median of these 10 000 k value combinations which account for the error on the predictor variables was then used to form one equation per location. The MATLAB code used to generate the boot resamples is shown in Code 1.

46

Code 1. MATLAB code used to generate the boot resamples and obtain the fitting coefficients (k_1, k_2 and k_3) using Equation 3.

```
%% Derive k values for sulfuric acid proxy concentration using Dada et al. 2020 equation
% fitCoeff(1) = k1
% fitCoeff(2) = k2
% fitCoeff(3) = k3

data = [CS, SO2, O3, Alkene, GlobRad]; %CS in s-1, SO2,O3,Alkene in cm-3, GlobRad in W/m2
H2SO4; %measured sulfuric acid in cm-3

% Create the fitting function according to Equation 3

Y_fit = @(fitCoeff,data) (-1).*(data(:,1)./(2*fitCoeff(3))) + ...
    sqrt((data(:,1)./(2*fitCoeff(3))).^2 + data(:,2)./(fitCoeff(3)).*...
    (fitCoeff(1).*data(:,5) + fitCoeff(2).*data(:,3).*data(:,4)));


% Obtain the fitting coefficients were obtained by minimizing the sum of the squared logarithm
%of the ratio between the proxy values and measured sulphuric acid concentration

sum_squared_error = @(fit_coeff) sum((log10(H2SO4 ./ (Y_fit(fit_coeff,data)))).^2);

%introduce bootstrap resampling

fit_index = 10000; %number of bootstrap resampling

[~,bootsam] = bootstrp(fit_index,sum_squared_error,data); %bootstrap resampling

%introduce uncertainty estimates on the measured predictor varianbles
%create an array of random floating-point numbers that are drawn from a
%uniform distribution in the open interval between the lower and upper bound of accuracy

% 20% uncertainty on condensation sink
a = log10(1/1.2); %lower bound accuracy
b = log10(1.2); %upper bound accuracy
r_CS = 10.^((b-a).*rand(fit_index,1)+a);

% factor of 2 uncertainty on H2SO4 measurement
a = log10(0.5);
b = log10(2);
r_SA = 10.^((b-a).*rand(fit_index,1) + a);

% 10% uncertainty on trace gases and global radiation
a = log10(1/1.1);
b =log10(1.1);
r_SO2 = 10.^((b-a).*rand(fit_index,1) + a); %SO2

a = log10(1/1.1);
b =log10(1.1);
r_O3 = 10.^((b-a).*rand(fit_index,1) + a); %O3

a = log10(1/1.1);
b =log10(1.1);
r_MT = 10.^((b-a).*rand(fit_index,1) + a); %Alkenes

a = log10(1/1.1);
b =log10(1.1);
r_GR = 10.^((b-a).*rand(fit_index,1) + a); %GlobRadiation

%
k_all=[];
for i =1:fit_index

    %create bootstrapped data disturbed with uncertainty on predictor variables
data_boot = [data(bootsam(:,i),1)*r_CS(i),data(bootsam(:,i),2)*r_SO2(i),...
    data(bootsam(:,i),3)*r_O3(i), data(bootsam(:,i),4)*r_MT(i),...
    data(bootsam(:,i),5)*r_GR(i)];
H2SO4_boot=H2SO4(bootsam(:,i),:)*r_SA(i);

% Obtain the fitting coefficients for the bootstrap resamples

sum_squared_error = @(fit_coeff) sum((log10(H2SO4_boot ./ (Y_fit(fit_coeff,data_boot)))).^2);

% Assume initial values for the fitting parameters:
k0 = [1e-8, 1e-27,1e-9];

% Use built-in MATLAB function fminsearch to find the fitting parameters;
% the best fit parameters are in output into variable k:
[k, SSE] = fminsearch (sum_squared_error, k0, options);

k_all = [k_all;k(:,:)];
end
```

Table S 1 Summary of measurement locations and instrumentation used for deriving the H_2SO_4 proxy (training data sets).

Location	Type	Measurement Period	Particle size distribution instrument	Trace Gases	Radiation
Hyytiälä, Finland	Boreal	August 18, 2016 to ~~April 16~~December 31, ~~2017~~ 2016 and March 8, 2018 to February 28, 2019	Twin ~~–~~ DMPS (Ground level).	SO_2 and O_3 are monitored using two Thermo Environmental Instruments (models 43i-TLE, 49i, respectively), at 16.8 m above ground level.~~.~~	[1]Global radiation was measured with Middleton solar SK08 pyranometer until August 24, 2017 and after that with Middleton solar EQ08-S pyranometer. at 16.8 m.
Agia Marina, Cyprus[2]	Rural background	February 22 and March 3, 2018	2-20 nm using Airel NAIS and 20-800 nm using TSI SMPS	SO_2 and O_3 are monitored using Ecotech Instruments (9850 and 9810, respectively)	Campbell Scientific weather station
Budapest, Hungary	Urban	March 21 and ~~May~~ April 17, 2018	6-1000 nm using flow-switching type DMPS	SO_2 is measured using UV fluorescence (Ysselbach 43C)	Global radiation was measured by an SMP3 pyranometer (Kipp and Zonnen, The Netherlands)
Beijing, China	MegaCity	~~December 1, 2018 – January 31, 2019~~March 15, 2019 – June 15, 2019	3 – 800 nm PSD system ~12 m above ground level.	SO_2 and O_3 are monitored using two Thermo Environmental Instruments (models 43i-TLE, 49i, respectively), ~12 m above ground..	[3]Global radiation was measured using CMP11 pyranometer (Kipp and Zonnen, Delft, Netherlands) at ~ 15 m above ground level.

[1] UVB radiation was measured with Solar SL 501A pyranometer.
[2] All variables are measured at the same height.
[3] UVB radiation was measured using a UVS-B-T radiometer (Kipp and Zonnen, Delft, Netherlands).

Table S 2 Summary of measurement locations and instrumentation used for verifying the predictive power of the derived proxies (testing data sets).

Location	Type	Measurement Period	Particle size distribution instrument	Trace Gases	Radiation
Hyytiälä, Finland	Boreal	January 1, 2017 – June 5, 2017	Twin –– DMPS(Ground level).	SO_2 and O_3 are monitored using two Thermo Environmental Instruments (models 43i-TLE, 49i, respectively).	Global radiation was measured with Middleton solar EQ08-S pyranometer.
Helsinki, Finland	Semi-urban	July 1, 2019 – July 16, 2019	Twin DMPS at ground level	SO_2 was measured using UV-flurescence (Horiba APSA 360) at 31 m above ground	Global radiation was monitored Kipp and Zonen CNR1 at 31 m above ground level
Beijing, China	MegaCity	September 8, 2019 – October 15, 2019	3 – 800 nm PSD system ~12 m above ground	SO_2 and O_3 are monitored using two Thermo Environmental Instruments (models 43i-TLE, 49i, respectively) ~ 12 m above ground..	Global radiation was measured using CMP11 pyranometer (Kipp and Zonnen, Delft, Netherlands) at ~ 15 m above ground level.
Kilpilahti, Finland	Industrial Area	June 07, 2012 – June 29, 2012	6 to 1000 nm DMPS.	SO_2 was monitored usingThermo Scientific ™ Model 43i SO_2 Analyser	Acquired from SMEAR III station.

Table S 3 Summary of basic statistics of measurements of condensation sink, trace gases and global radiation at all locations and time periods included in this study. For Hyytiälä, Beijing and Kilpilahti we use all day time window, for Agia Marina, Budapest and Helsinki we use daytime statistics (GlobRad > 50 W/m^2).

Location		**Hyytiälä, Finland**	**Hyytiälä, Finland**	**Agia Marina, Cyprus**	**Helsinki, Finland**	**Budapest Hungary**	**Beijing, China**	**Beijing, China**	**Kilpilahti, Finland**
Type		Boreal	Boreal	Rural	Semi-urban	Urban	MegaCity	MegaCity	Industrial Area
Measurement Period		August 18 - December 31, 2016 March 8 - February 28, 2019	January 1, 2017 – June 5, 2017	February 22 - March 3, 2018	July 1 – July 16, 2019	March 21 - April 17, 2018	March 15, 2019 – June 15, 2019	September 8, 2019 – October 15, 2019	June 07, 2012 – June 29, 2012
CS (10^{-3} s^{-1})	mean	4.48	2.88	4.43	3.38	11.74	24.20	23.22	5.25
	median	3.83	2.18	3.63	3.13	10.92	22.83	22.60	4.91
	5th percentile	0.85	0.74	1.37	1.25	5.03	7.60	5.14	2.61
	95th percentile	12.43	8.78	9.58	6.47	21.52	44.58	44.34	8.81
	sd	3.89	2.42	2.55	1.60	5.37	11.86	11.82	2.11
SO_2 (10^{10} cm^{-3})	mean	0.31	0.30	0.70	1.30	6.02	4.70	2.43	6.65
	median	0.12	0.16	0.46	0.87	5.45	3.49	1.35	2.98
	5th percentile	0.03	0.01	0.17	0.13	3.35	0.26	0.13	0.99
	95th percentile	1.24	1.01	1.96	2.19	12.42	13.71	8.47	26.00
	sd	0.54	0.47	0.65	3.19	2.54	4.59	3.56	11.46
O_3 (10^{10} cm^{-3})	mean	83.59	95.08				105.63	116.10	161.36
	median	80.27	97.10				95.66	102.53	178.15
	5th percentile	41.09	65.42				5.23	3.24	24.81
	95th percentile	134.85	118.42				238.26	260.97	234.37
	sd	28.52	16.80				72.22	80.99	62.92
Alkene (10^{10} cm^{-3})	mean	0.92	0.32				14.33	11.98	2.27
	median	0.39	0.15				12.29	11.91	0.72
	5th percentile	0.05	0.02				1.91	2.55	0.11

	95th percentile	3.54	0.85				34.40	19.51	10.20
	sd	2.03	0.98				9.68	4.96	3.38
Global Radiation ($W.m^{-2}$)	mean	149.25	93.06	283.71	353.67	322.90	243.72	221.27	307.86
	median	47.53	23.17	272.48	270.60	300.56	54.27	52.97	252.64
	5th percentile	0.47	0.36	67.92	61.59	70.64	0.02	0.02	0.06
	95th percentile	636.60	378.50	548.90	837.27	697.42	840.95	730.83	768.84
	sd	205.18	137.32	155.33	254.08	200.36	308.33	273.10	280.05
H_2SO_4 (10^6 cm^{-3})	mean	0.73	0.55	2.76	3.82	1.54	2.94	3.45	10.59
	median	0.28	0.18	1.81	2.55	1.02	1.61	2.00	3.19
	5th percentile	0.02	0.02	0.17	0.41	0.23	0.37	0.37	0.19
	95th percentile	2.55	2.01	8.22	11.71	4.76	8.63	10.98	37.08
	sd	1.40	1.06	3.06	4.57	1.77	3.00	3.74	28.25

65

Table S 4 Statistical parameters included in deriving the Aikake Information Criterion. Equation number refers to the number in the main text, N is the sample size (number of points), X is the number of coefficients (number of k values) and SSE is the sum of squared estimate of errors. AIC is calculated as AIC = 2X + N ln(SSE). The quantity exp((AIC$_{min}$ − AIC$_i$)/2) describes the probability that the ith model minimizes the information loss. For example, Equation 5 in Hyytiälä is 5.62E-8 times as probable as the Equation 6 to minimize the information loss.

Hyytiälä Eq. 9	Equation number	6	5	4	2
	number of coefficients	3	2	2	1
	N	1860	1860	1860	1860
	R	0.84	0.74	0.82	0.70
	Slope	0.80	0.78	0.96	1.84
	SSE	1.89E+02	3.00E+02	2.88E+02	1.17E+03
	AIC	4.24E+03	4.61E+03	4.58E+03	5.71E+03
	exp((AIC$_{min}$ − AIC$_i$)/2)	1	5.62E-81	5.09E-74	0
Cyprus Eq. 10	Equation number	6	5	4	2
	number of coefficients	3	2	2	1
	N		96		96
	R		0.88		0.80
	Slope		0.53		0.67
	SSE		2.02		5.22
	AIC		33.30		69.86
	exp((AIC$_{min}$ − AIC$_i$)/2)		1		1.15E-08
Budapest Eq. 11	Equation number	6	5	4	2
	number of coefficients	3	2	2	1
	N		263		263
	R		0.59		0.49
	Slope		0.47		0.95
	SSE		10.73		30.10
	AIC		275.06		389.85
	exp((AIC$_{min}$ − AIC$_i$)/2)		1		1.19E-25
Beijing Eq. 12	Equation number	6	5	4	2
	number of coefficients	3	2	2	1
	n	877	877	877	877
	R	0.72	0.89	0.70	0.90
	Slope	1.69	3.16	2.11	5.23
	SSE	189.72	318.04	275.05	769.09
	AIC	2003.90	2198.67	2143.37	2532.00
	exp((AIC$_{min}$ − AIC$_i$)/2)	1	2.57E-85	2.69E-61	4.4E-230

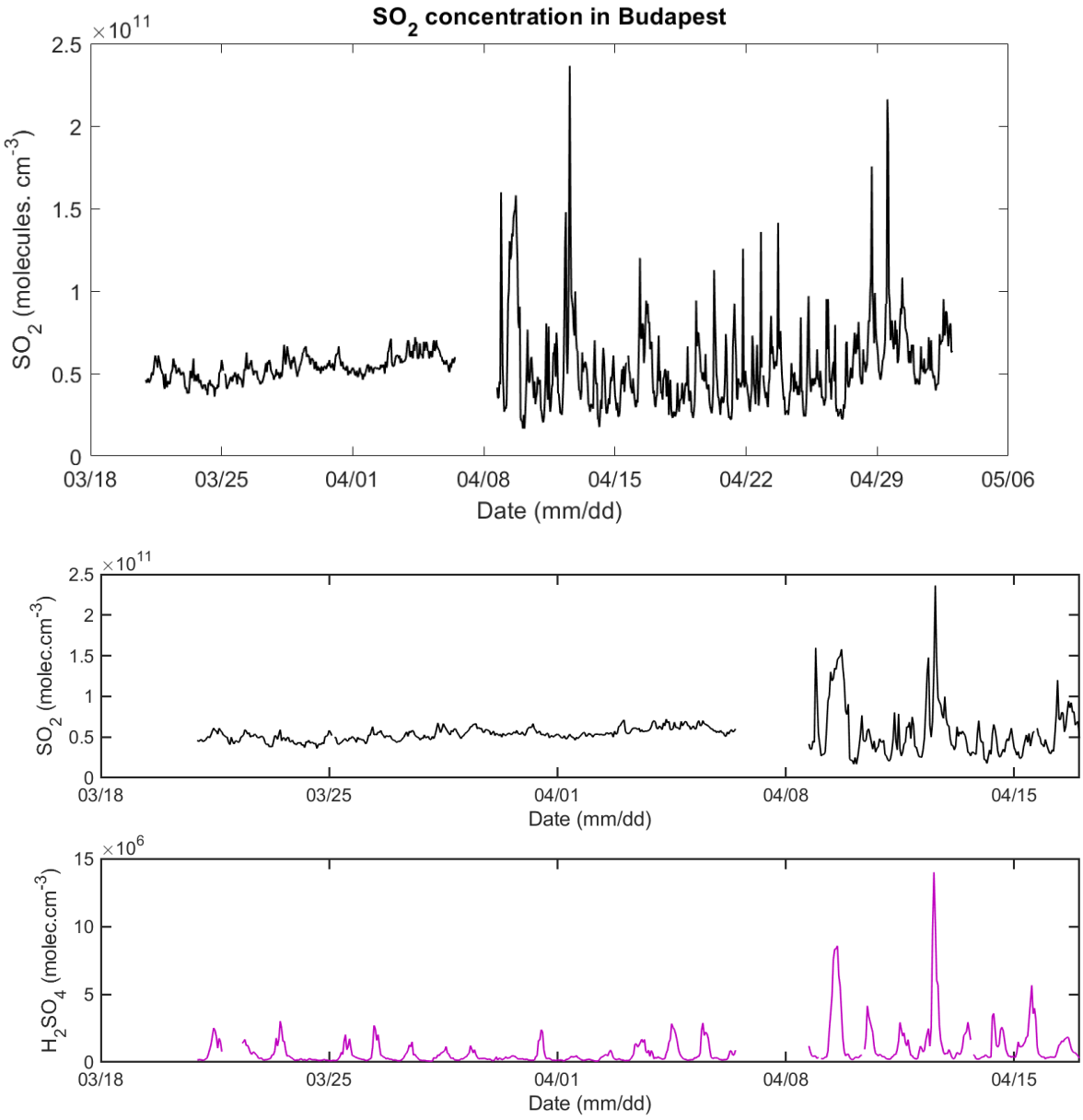
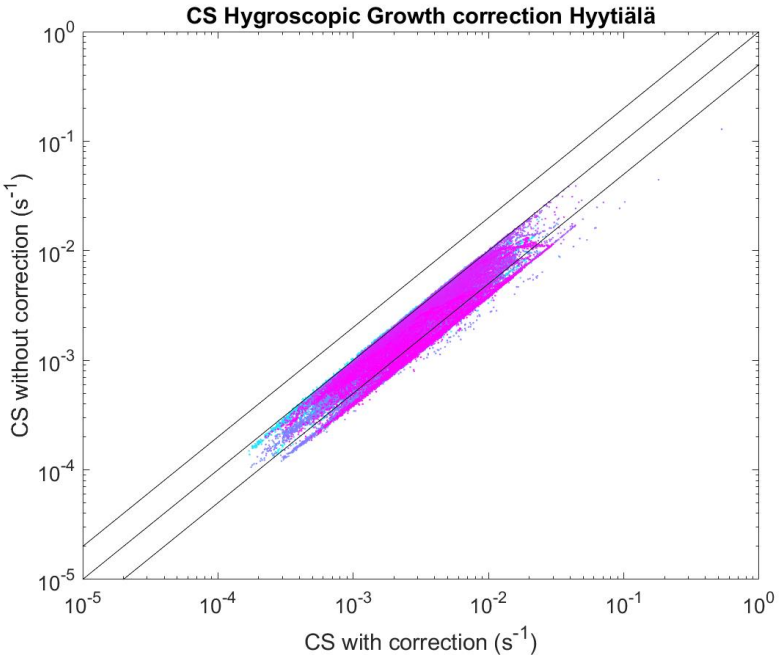
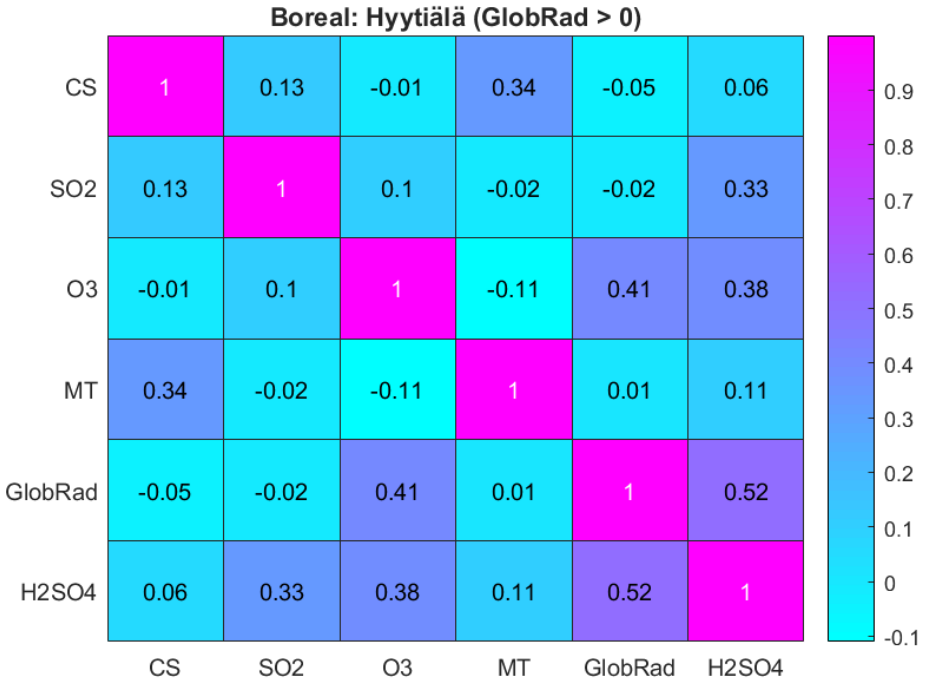


Figure S 1 SO_2 and measured H_2SO_4 concentrations in Budapest showing the change in concentration due to changes in meteorology mid-campaign.



~~*Figure S 2 Effect of hygroscopic growth correction on condensation sink calculation in the boreal forest. Figure S 3*~~

~~Figure S 2~~



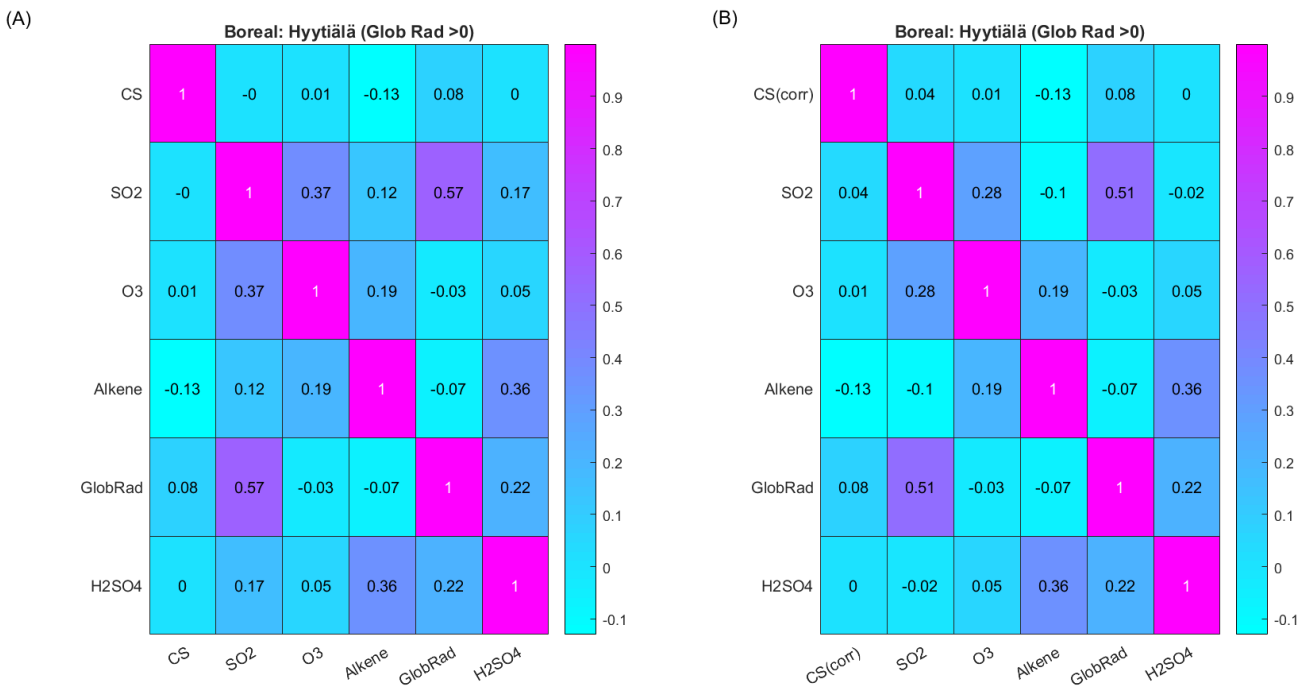
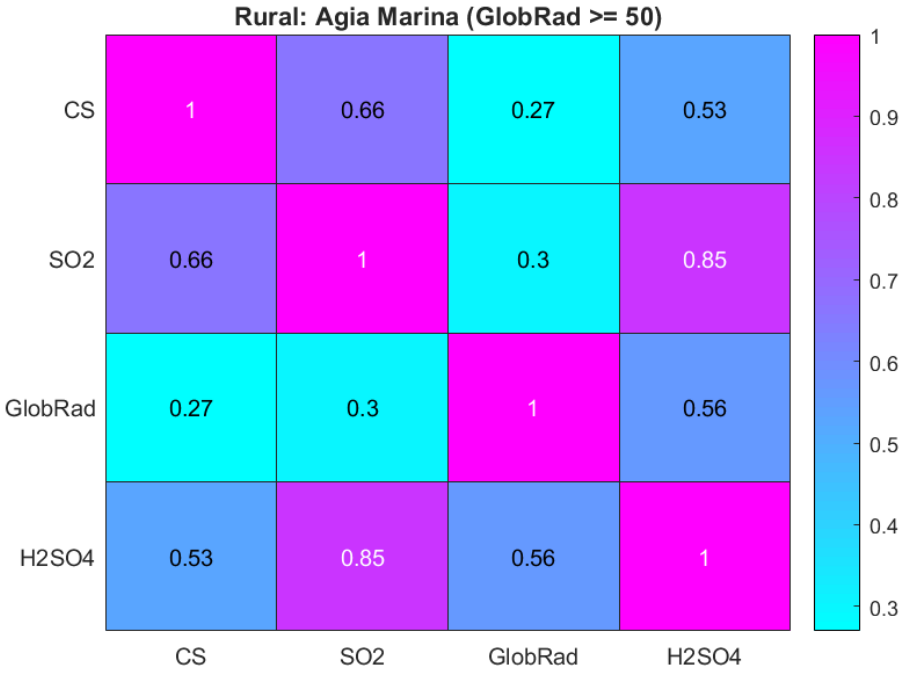


Figure S 2 ~~Pearson's~~Spearman's correlation coefficients matrix between variables involved in H_2SO_4 formation and loss at the Hyytiälä station (Global Radiation > 0 W/m^2). CS represents condensation sink in s^{-1}. SO_2, O_3 and MT (monoterpenes) in molecules/cm^{-3}. GlobRad is global radiation in W/m^2. H_2SO_4 is measured sulphuric acid in molecules/cm^{-3}. The color bar represents the Spearman's correlation coefficient. In (A) the condensation sink is not corrected for hygroscopic growth, while in (B) the condensation sink is corrected for hygroscopic growth using the parametrization given by Laakso et al. (2004).



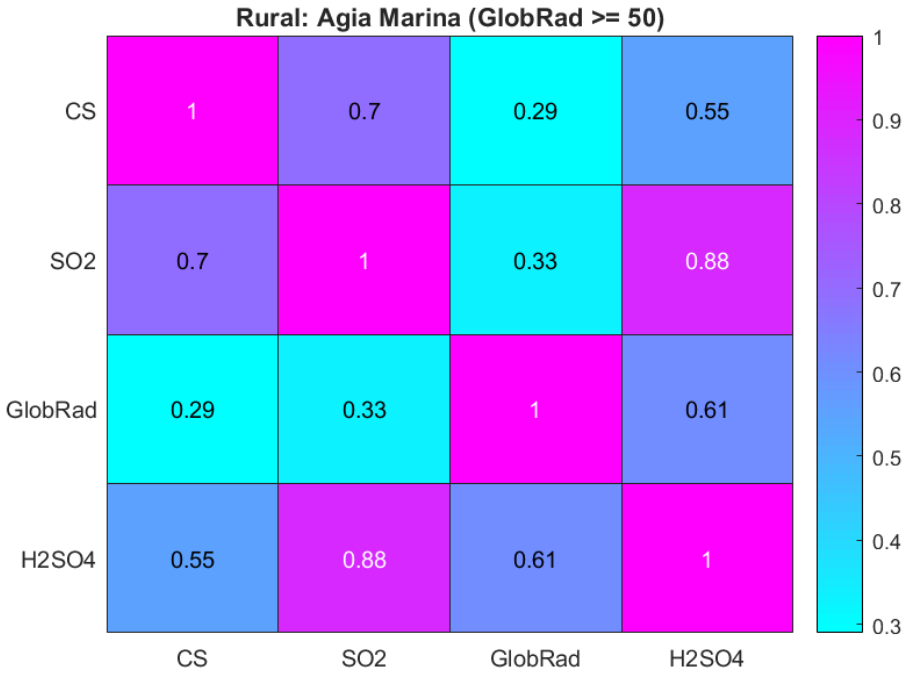
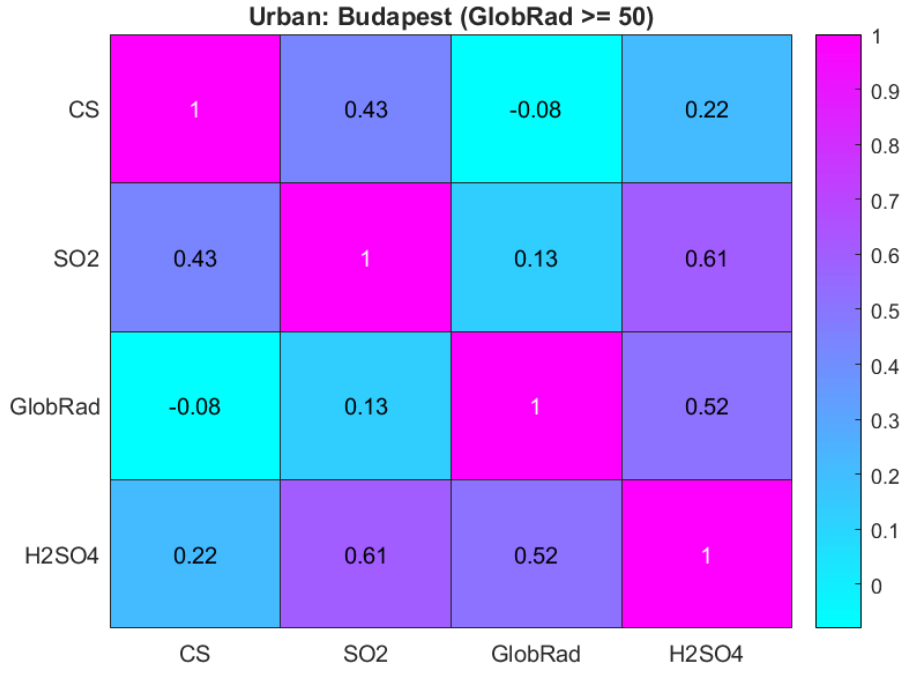


Figure S 3 ~~Pearson's~~ Spearman's correlation coefficients matrix of variables involved in H_2SO_4 formation and loss at the Agia Marina station (Global Radiation > 50 W/m^2). CS represents condensation sink in s^{-1}. SO_2 is in molecules/cm^{-3}. GlobRad is global radiation in W/m^2. H_2SO_4 is measured sulphuric acid in molecules/cm^{-3}. The color bar represents the Spearman's correlation coefficient.



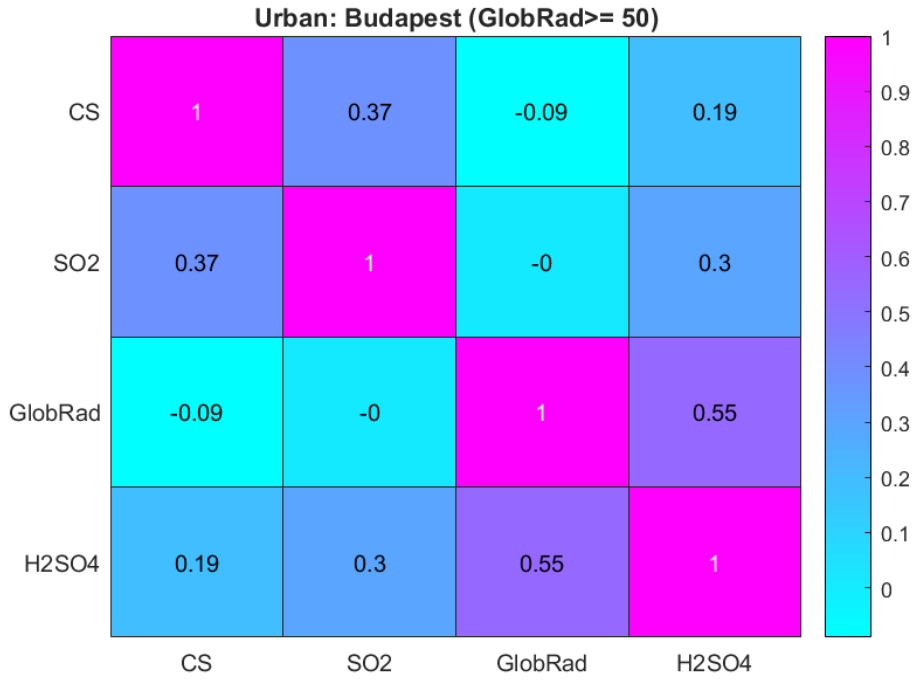
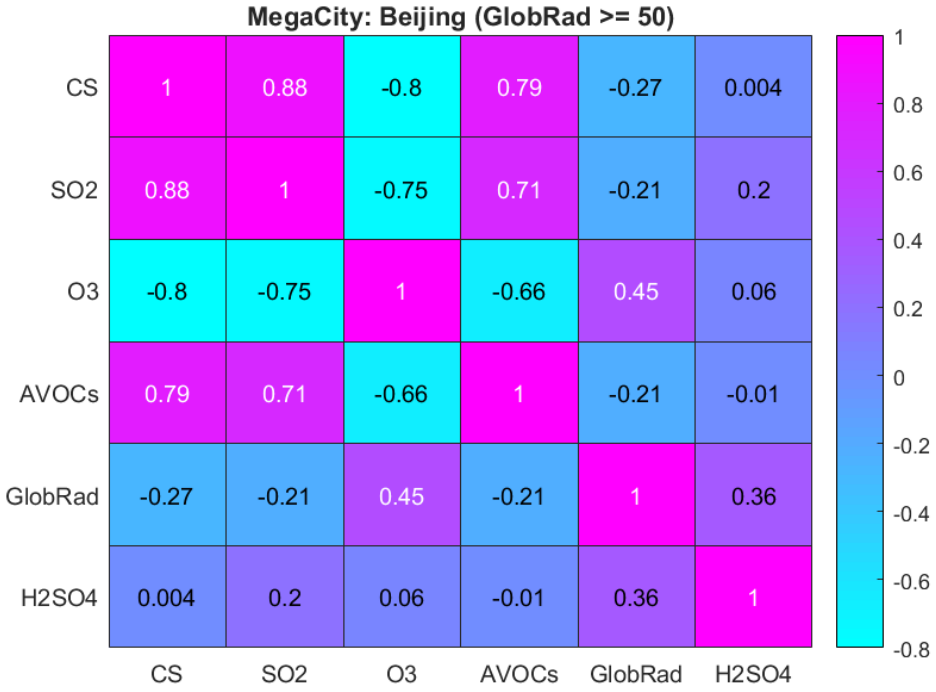


Figure S 4 ~~Pearson's~~Spearman's correlation coefficients matrix of variables involved in H_2SO_4 formation and loss at the Budapest station (Global Radiation > 50 W/m^2). CS represents condensation sink in s^{-1}. SO_2 in molecules/cm^{-3}. GlobRad is global radiation in W/m^2. H_2SO_4 is measured sulphuric acid in molecules/cm^{-3}. The color bar represents the Spearman's correlation coefficient.



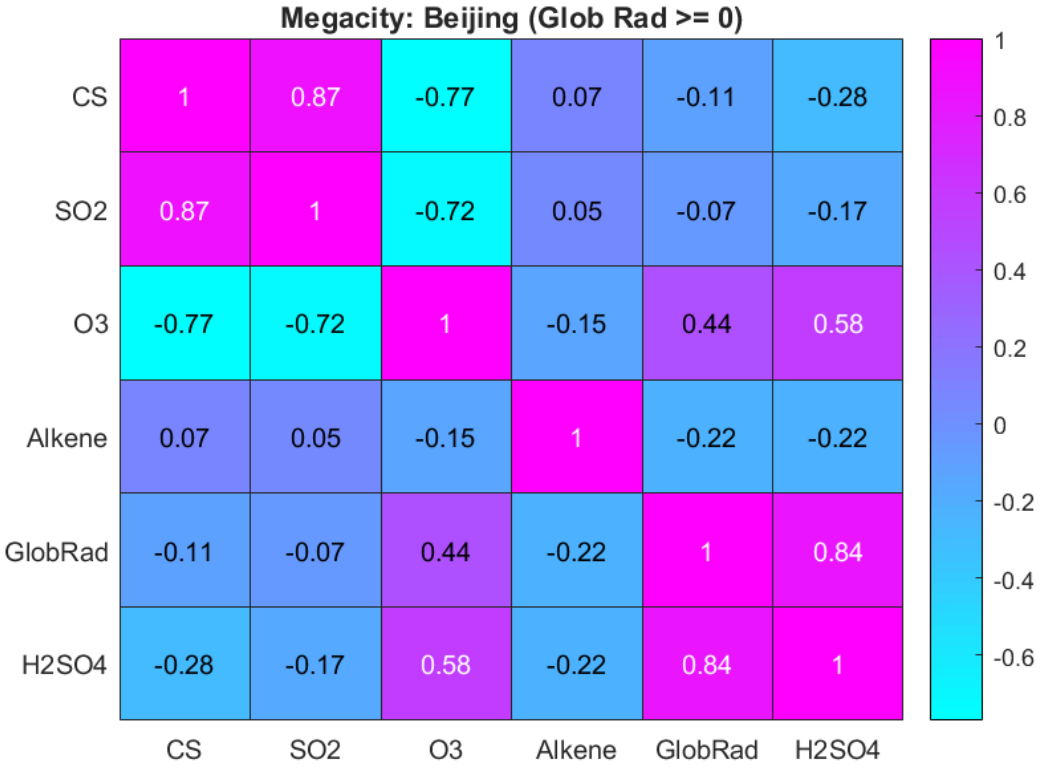
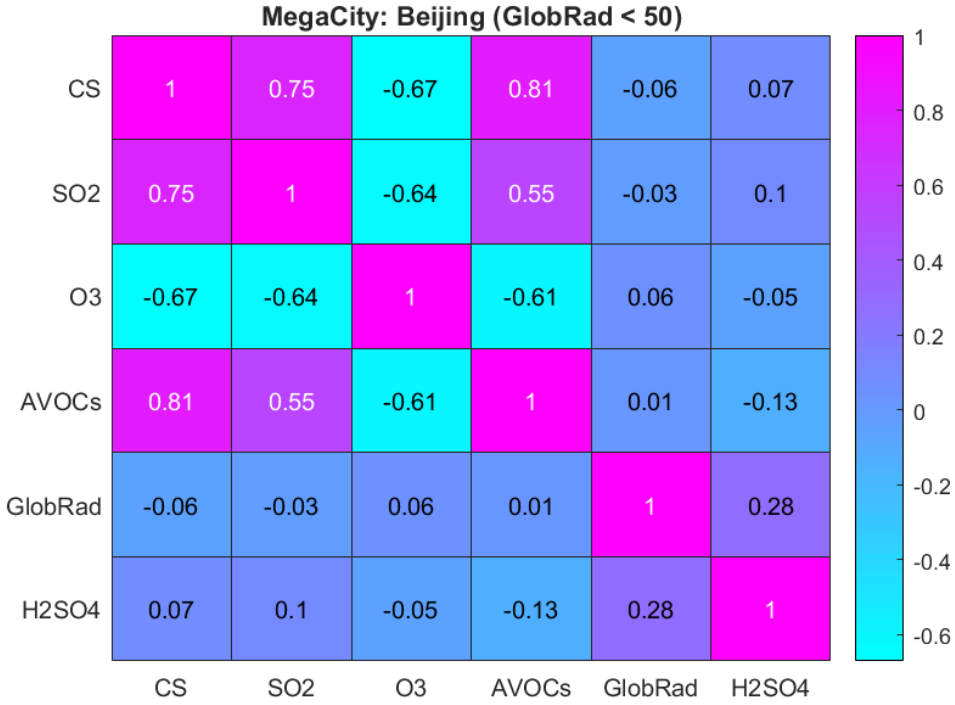


Figure S 5 ~~Pearson's~~Spearman's correlation coefficients matrix between variables involved in H_2SO_4 formation and loss at the Beijing station ~~during daytime (Global Radiation > 50 W/m^2)~~. CS represents condensation sink in s^{-1}. SO_2, O_3 and ~~AVOCs~~ Alkenes (Anthropogenic volatile organic compounds) in molecules/cm^{-3}. GlobRad is global radiation in W/m^2. H_2SO_4 is measured sulphuric acid in molecules/cm^{-3}. The color bar represents the Spearman's correlation coefficient.



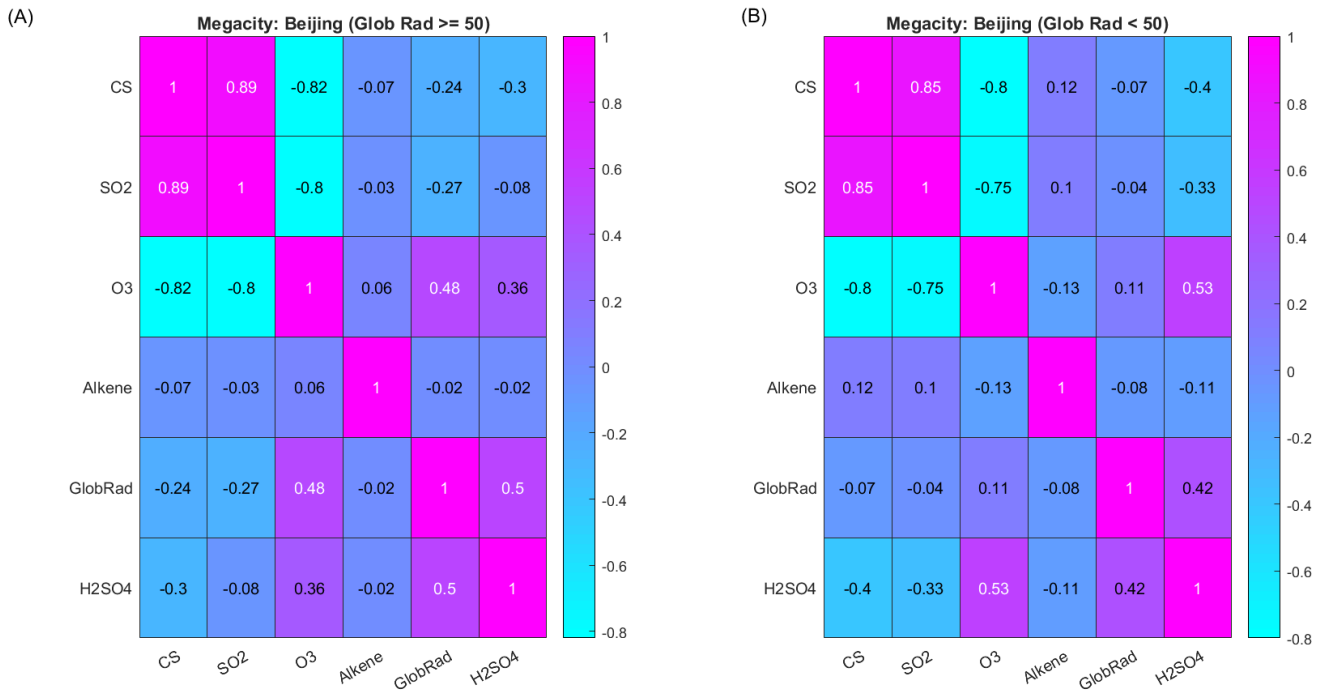


Figure S 6 ~~Pearson's~~Spearman's correlation coefficients matrix between variables involved in H_2SO_4 formation and loss at the Beijing station ~~during nighttime (Global Radiation < 50 W/m^2)~~. CS represents condensation sink in s^{-1}. SO_2, O_3 and ~~AVOCs~~ Alkenes (Anthropogenic volatile organic compounds) in molecules/cm^{-3}. GlobRad is global radiation in W/m^2. H_2SO_4 is measured sulphuric acid in molecules/cm^{-3}. The color bar represents the Spearman's correlation coefficient. In (A) the daytime correlation coefficients are shown (Global radiation >= 50 W/m^2) and in (B) the nighttime correlation coefficents are shown (Global radiation < 50 W/m^2).

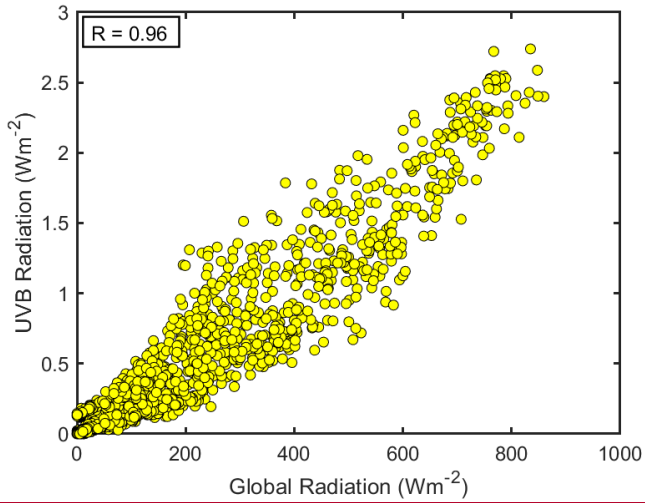


Figure S 7 Comparison between Global radiation and UVB in Hyytiälä. Hourly medians are shown. The total number of data points in the plot is 2306.

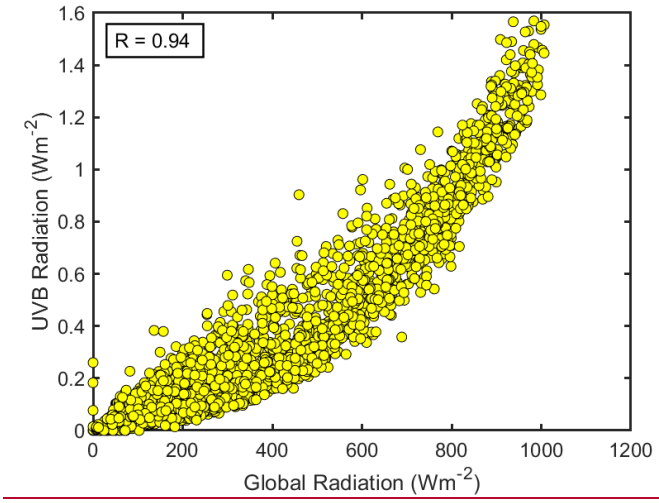


Figure S 8 Comparison between Global radiation and UVB in Beijing. Hourly medians are shown. The total number of data points in the plot is 7106.

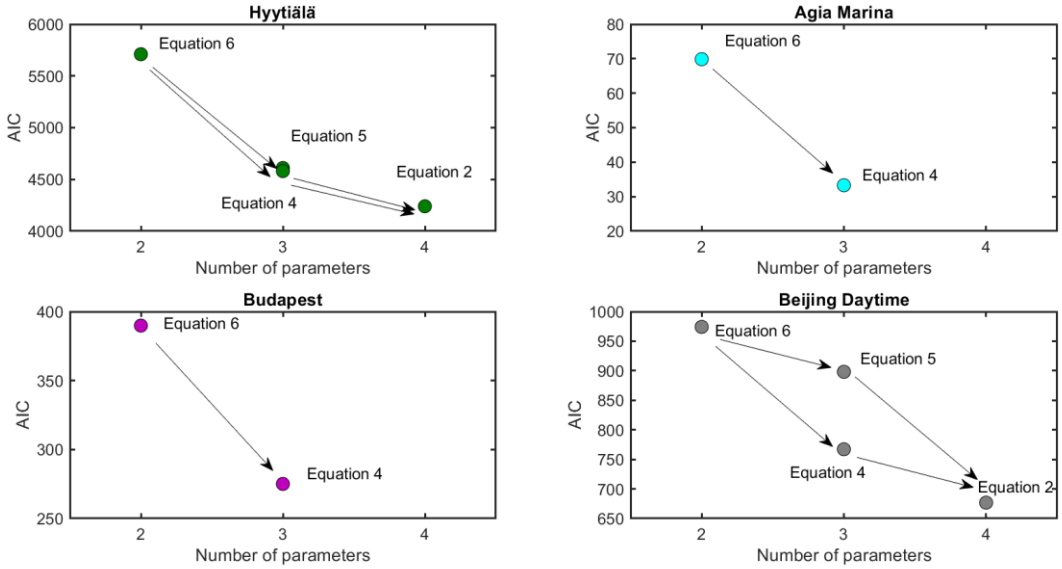


Figure S 9 Evaluation of the goodness of the fit using the Akaike information criterion (AIC) (McElreath, 2018). Number of parameters refers to the number of variables in each equation used. For example, Equation 2 uses four parameters which are the two sources (Radiation and sCI) and the two sinks (CS and cluster formation).

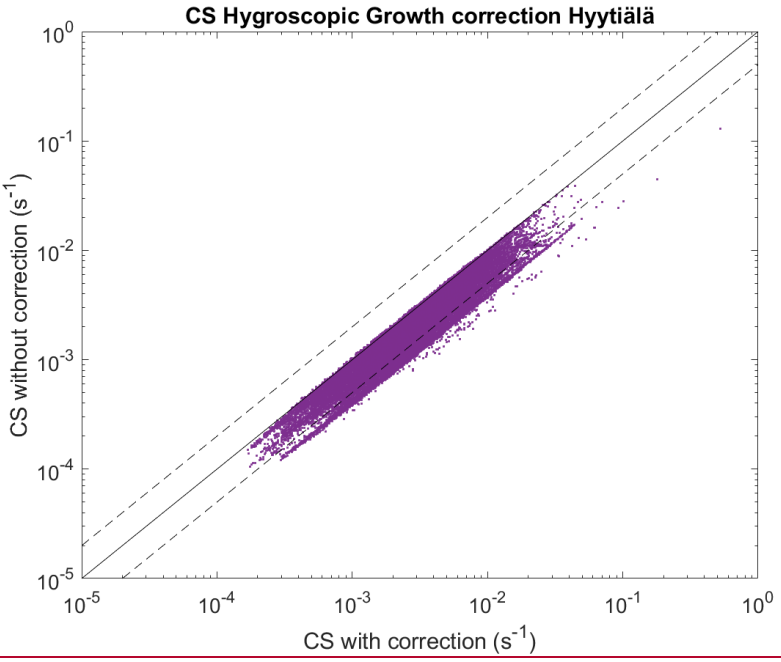


Figure S 10 Effect of hygroscopic growth correction on condensation sink calculation in the boreal forest. The solid line is the 1:1 line and the dashed lines are the 2:1 lines.

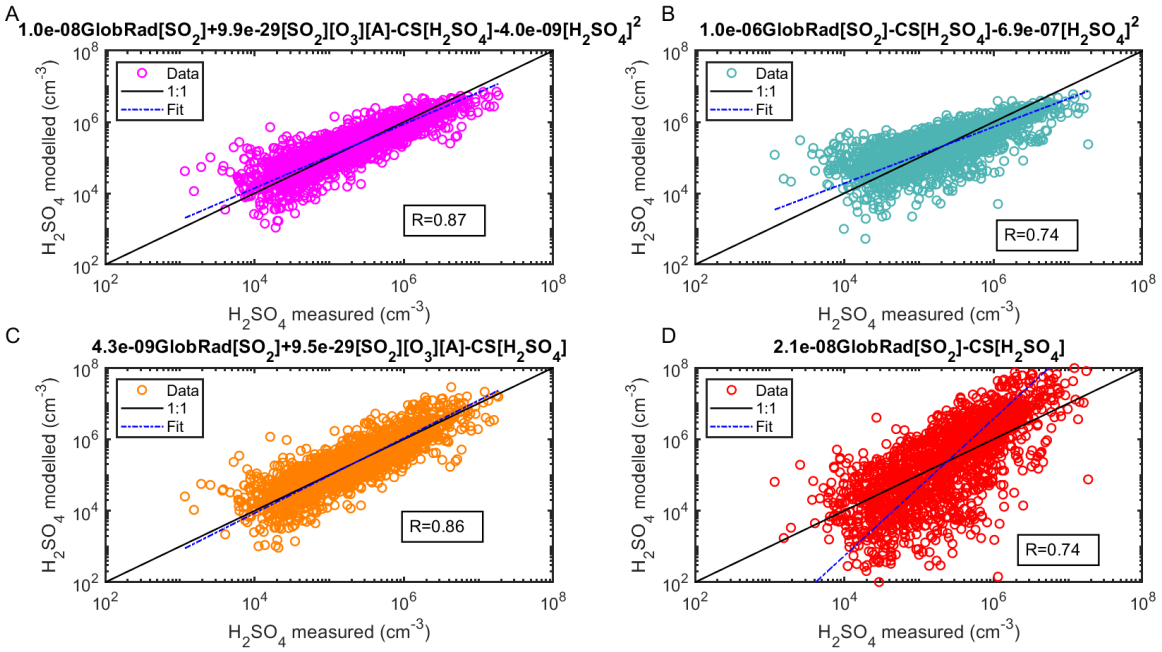


*Figure S 11 Sulphuric acid proxy concentration as a function of measured sulphuric acid. Observation at SMEAR II station, Hyytiälä Finland **with CS corrected for hygroscopic growth**. The observed concentrations are measured 2016-2019 using CI-APi-ToF and are 3-hour medians resulting in a total of 1594 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration(log(y)=a.log(x)+b).*

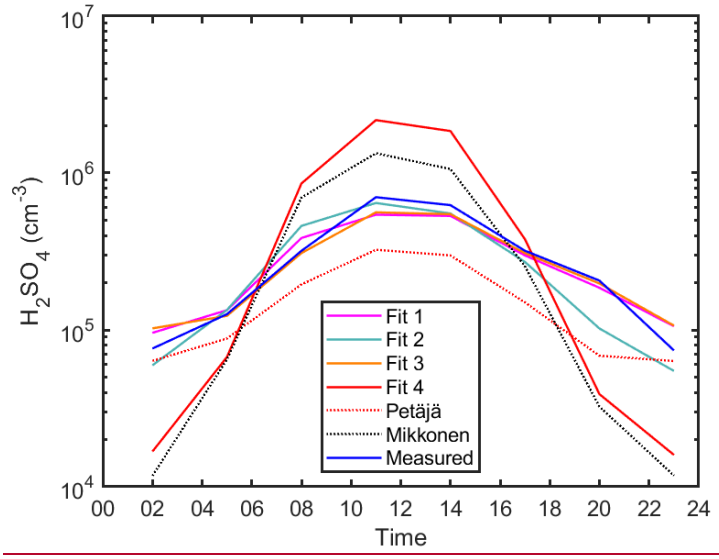
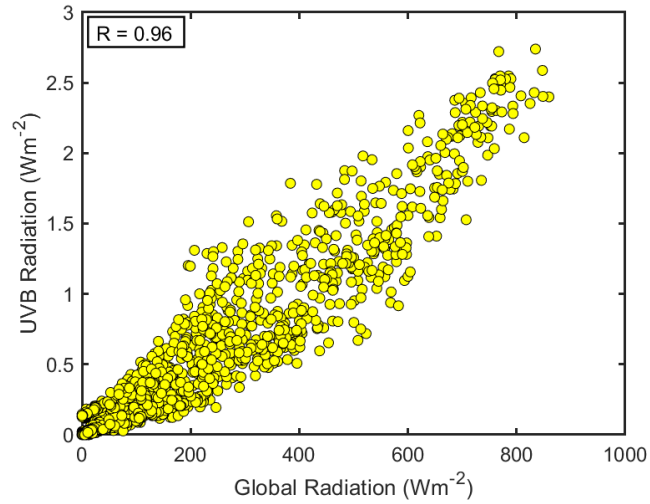
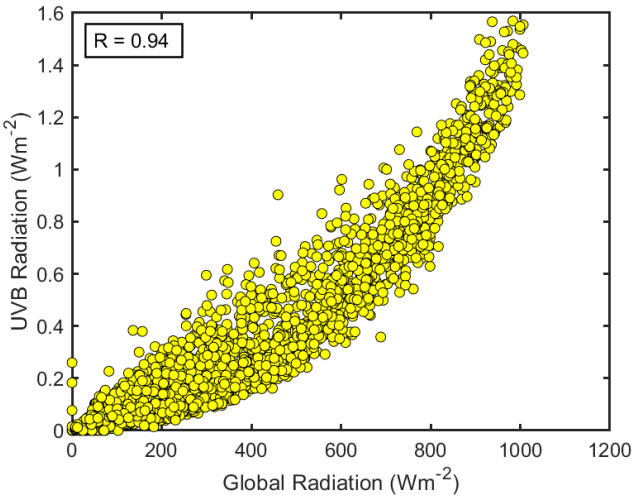


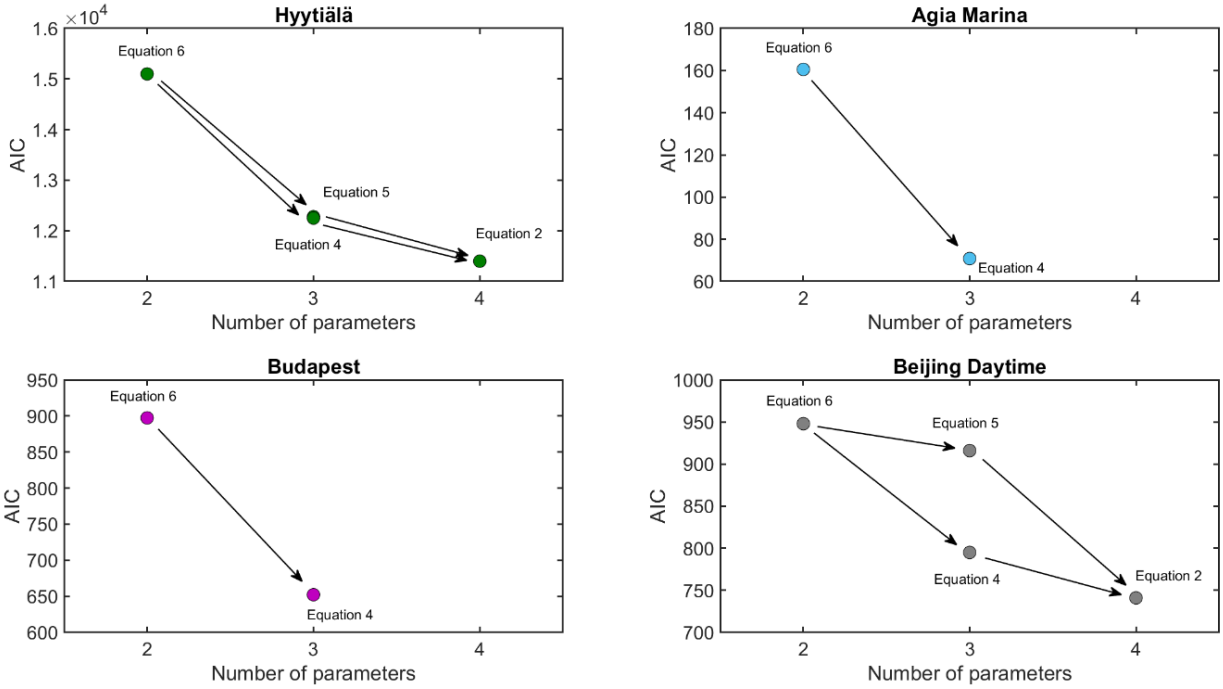
Figure S 12 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR II in Hyytiälä, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).



~~*Figure S 8 Comparison between Global radiation and UVB in Hyytiälä. Hourly medians are shown. The total number of data points in the plot is 2306.*~~



~~*Figure S 9 Comparison between Global radiation and UVB in Beijing. Hourly medians are shown. The total number of data points in the plot is 7106.*~~



~~*Figure S 10 Evaluation of the goodness of the fit using the Akaike information criterion (AIC) (McElreath, 2018). Number of parameters refers to the number of variables in each equation used. For example, Equation 2 uses four parameters which are the two sources (Radiation and sCI) and the two sinks (CS and cluster formation).*~~

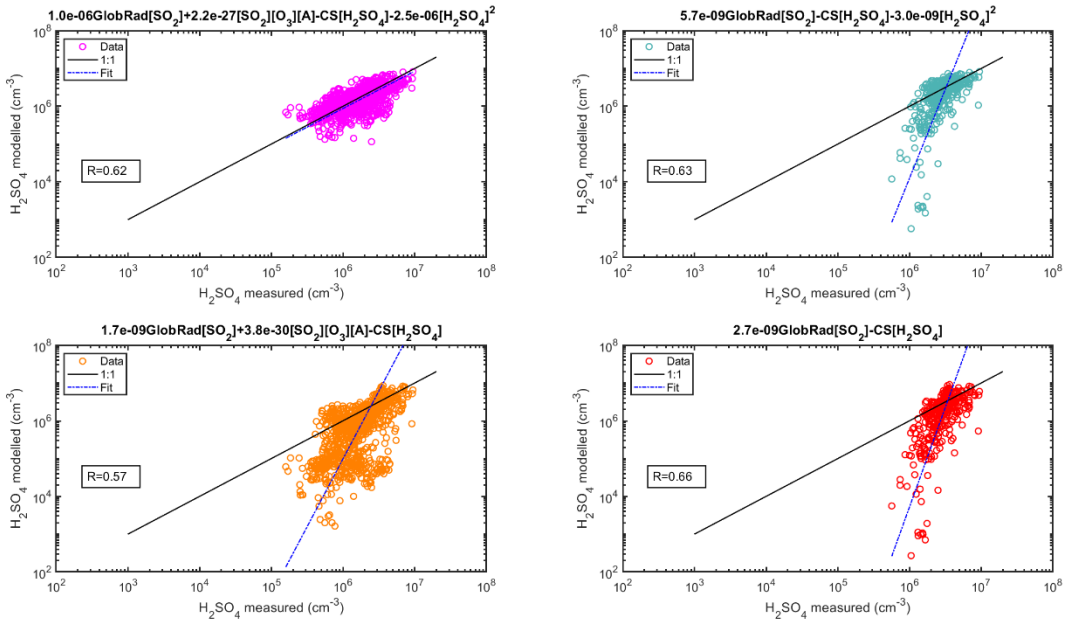


Figure S 11 Sulphuric acid proxy concentration as a function of measured sulphuric acid. observation at BUCT station, Beijing, China for day and nighttime combined. The observed concentrations are measured 2018-2019 using CI-APi-ToF and are 1-hour medians resulting in a total of 902 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Crigee Intermediates source and the cluster sink term (Equation 6).

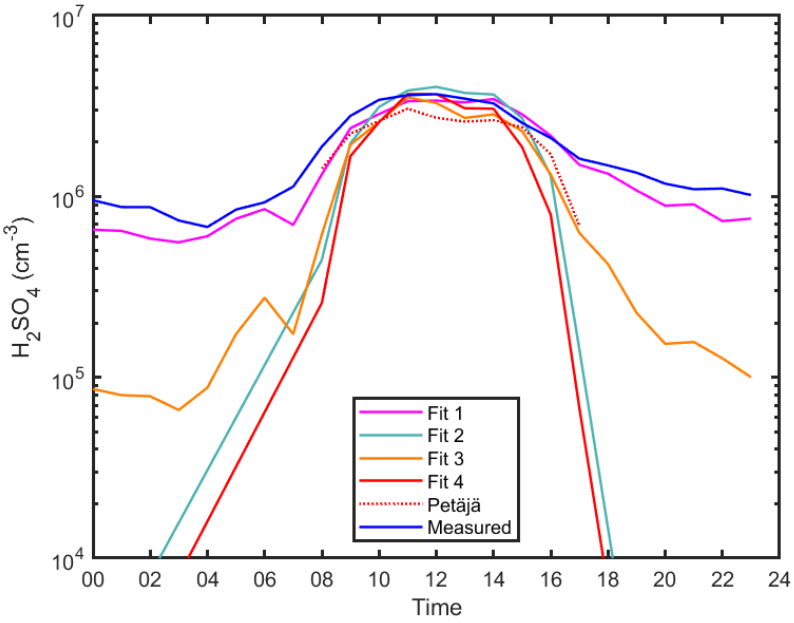
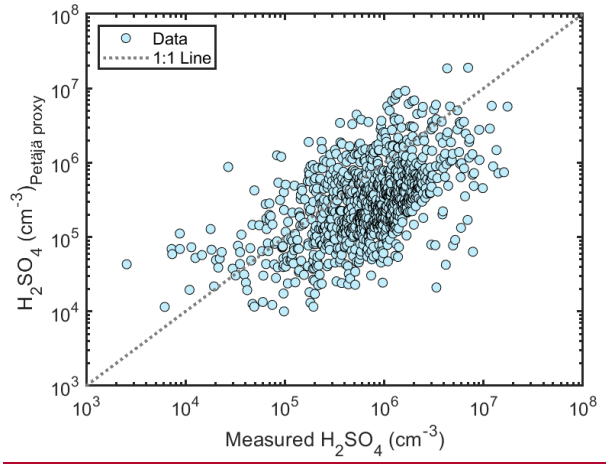
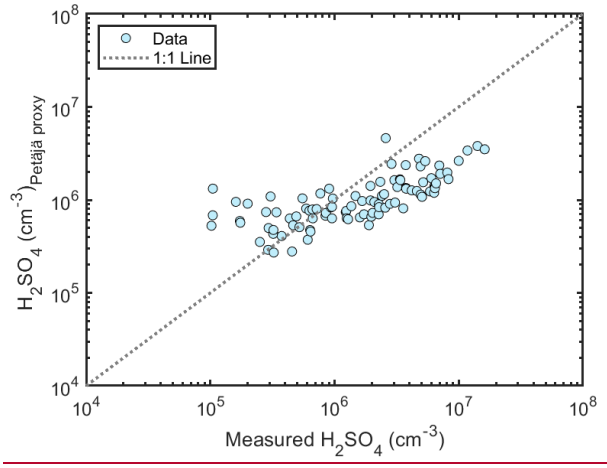


Figure S 12 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at Beijing China. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009.

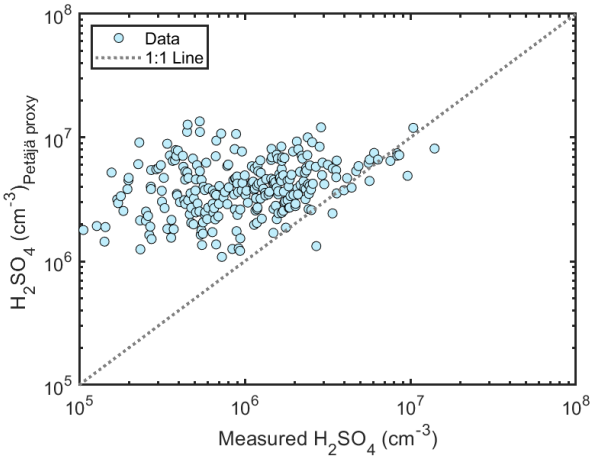
Hyytiälä



Agia Marina



Budapest



Beijing

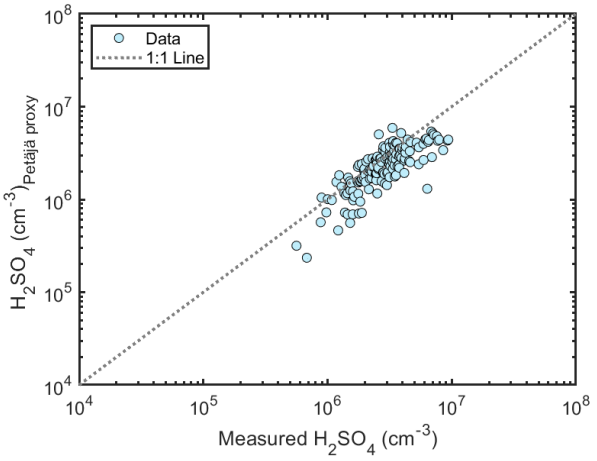
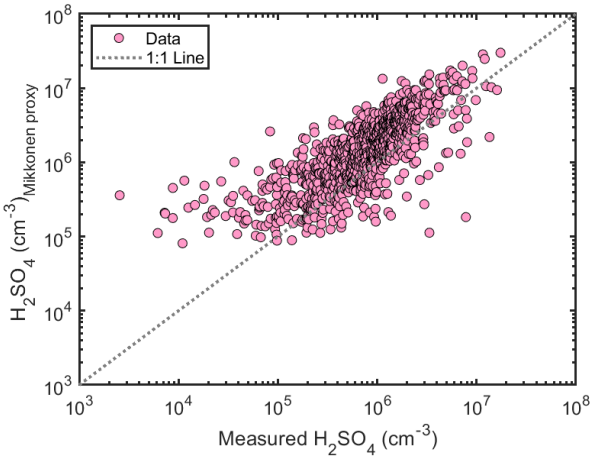
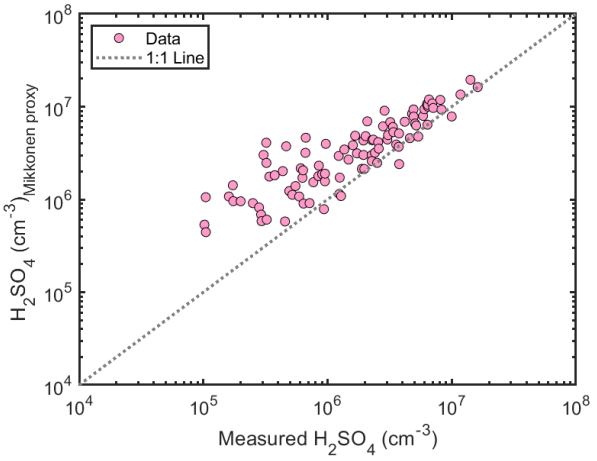


Figure S 13 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Petäjä et al. 2009 proxy at the 4 locations during daytime(GlobRad >= 50 W/m^2): Hyytiälä, Agia Marina, Budapest and Beijing.

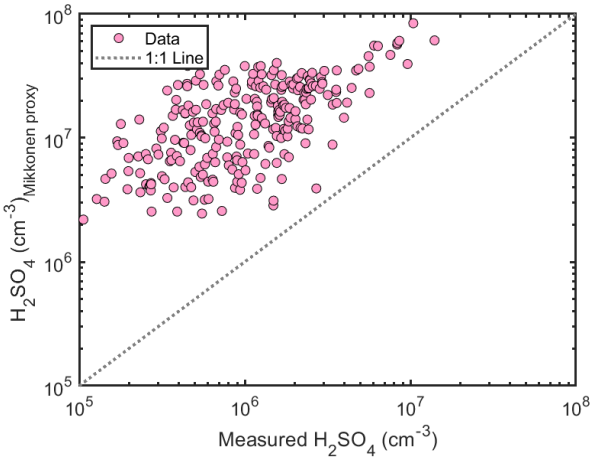
Hyytiälä



Agia Marina



Budapest



Beijing

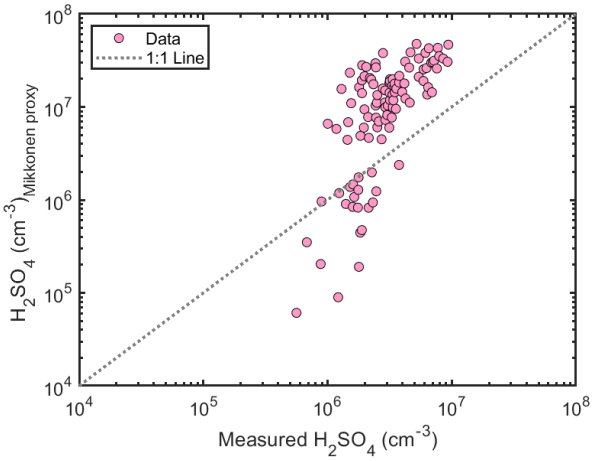


Figure S 14 Scatter plot showing the correlation between measured sulphuric acid and the sulphuric acid concentrations derived from the Mikkonen et al. 2011 proxy at the 4 locations during daytime (GlobRad >= 50 W/m^2): Hyytiälä, Agia Marina, Budapest and Beijing.

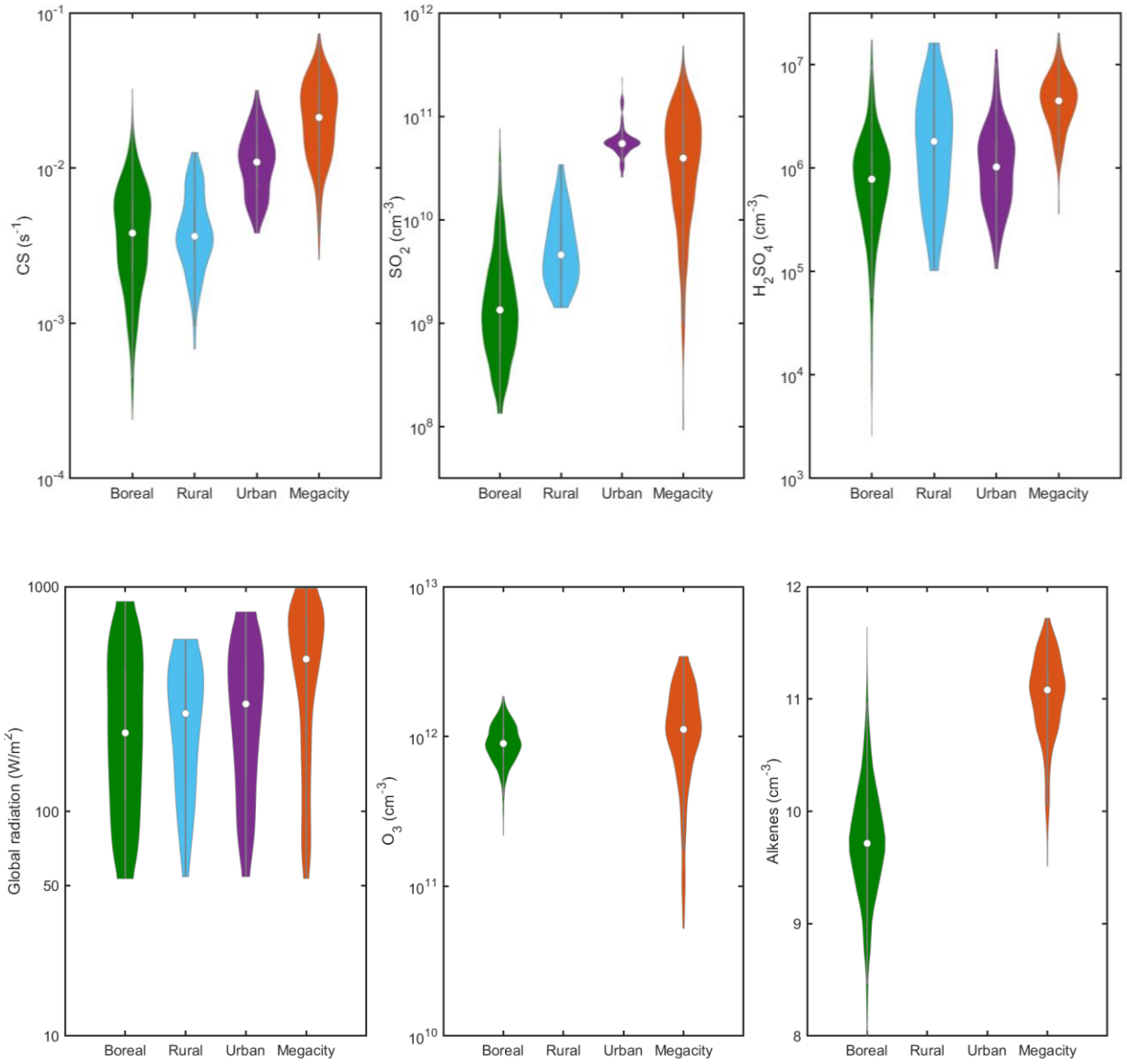


Figure S 15 Daytime data (GlobRad > 50 W/m^2) condensation sink, SO_2,GlobRad and H_2SO_4 concentrations in diffrerent environements. The concentrations are displayed as violin plots which are a combination of boxplot and a kernel distribution function on each side of the boxplots. The white circles define the median of the distribution and the edges on the inner grey boxes refer to the 25th and 75th percentiles respectively.

(A) *(B)*

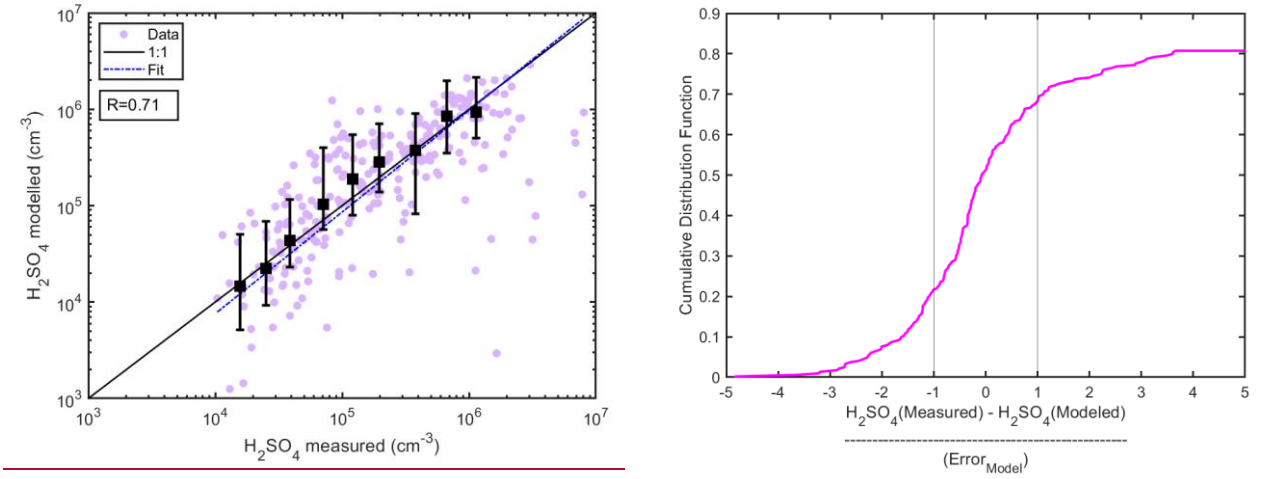


Figure S 16 (A) Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Hyytiälä SMEAR II station. The concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations. (B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 257 data points).

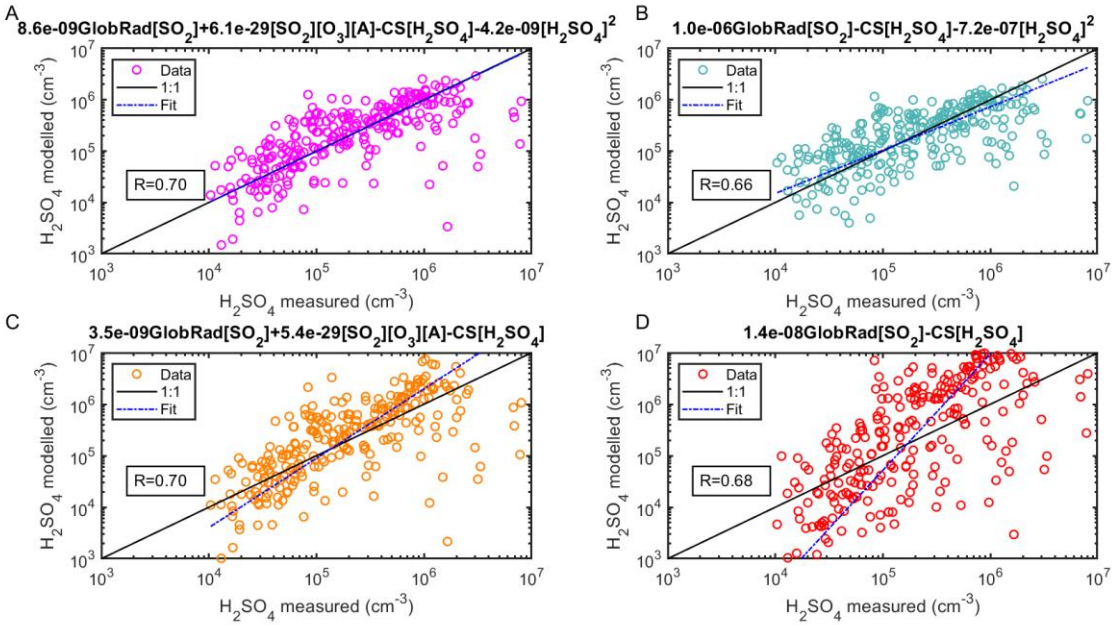


Figure S 17 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at SMEAR II station, Hyytiälä Finland using the four different combinations of source and sink terms. The concentrations shown are 3-hour medians coinciding with the alkene measurements every three hours resulting in a total of 257 data points. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the

cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration.

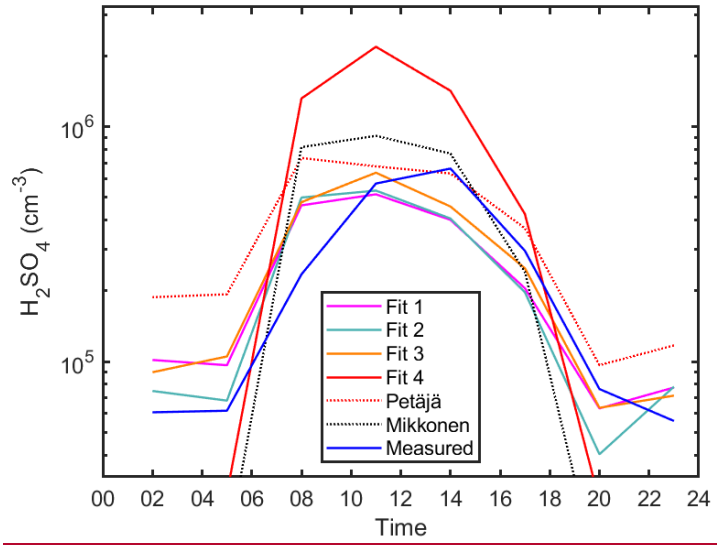
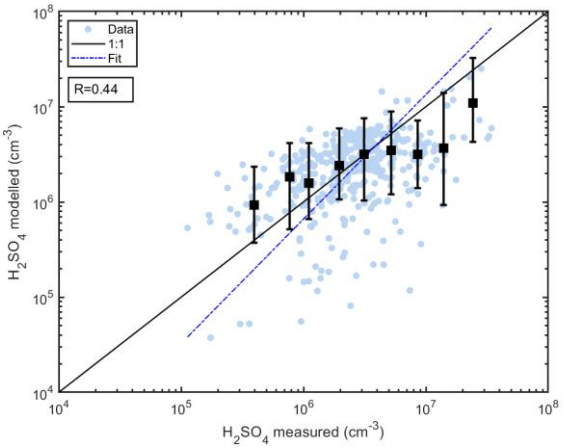


Figure S 18 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR II in Hyytiälä, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in (Petäjä et al., 2009)(Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

(A)



(B)

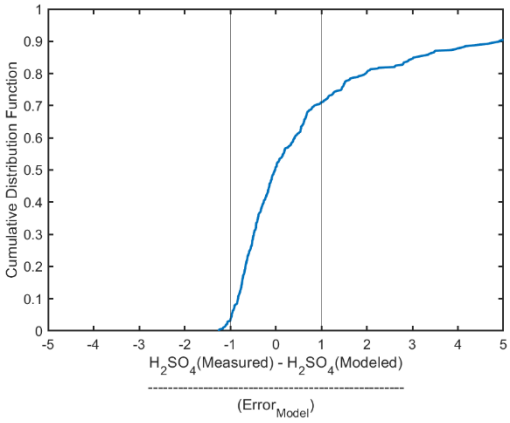


Figure S 19 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Helsinki SMEAR III station. The concentrations shown are 1-hour medians resulting in a total of 416 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations. (B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 416 data points).

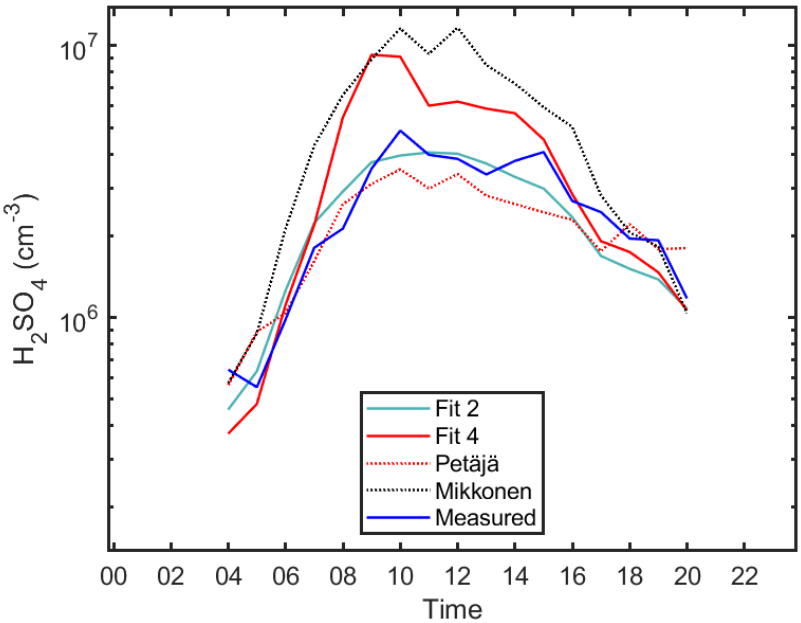


Figure S 20 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at SMEAR III in Helsinki, Finland. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

(A) (B)

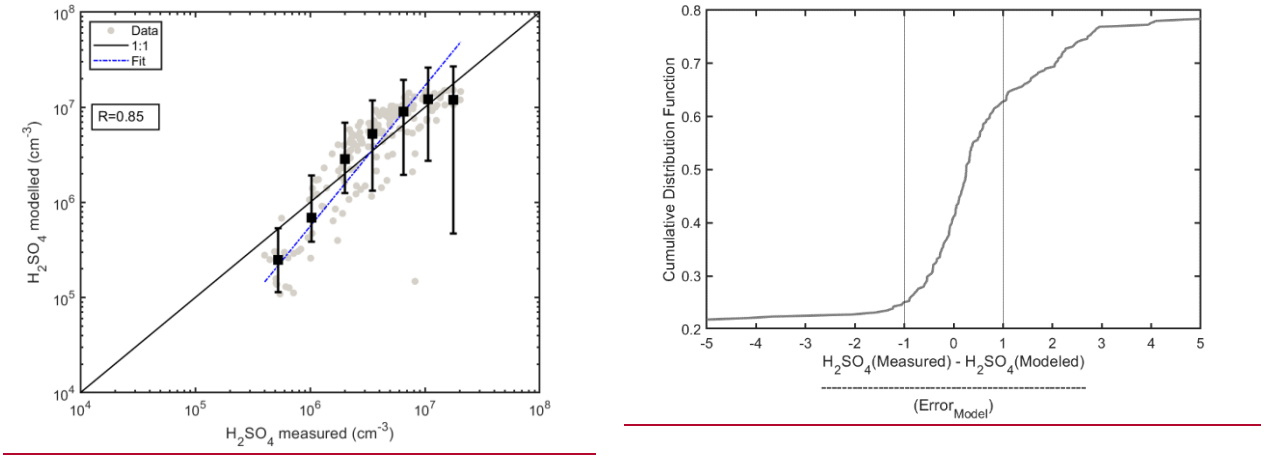


Figure S 21 Sulphuric acid concentrations modelled as a function of measured sulphuric acid in Beijing. The concentrations shown are 1-hour medians resulting in a total of 263 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the site. The gray data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whiskers correspond to 25th and 75th percentiles of the predicted concentrations.(B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 268 data points). H_2SO_4 concentration relative to the measured H_2SO_4 concentration (using 268 data points).

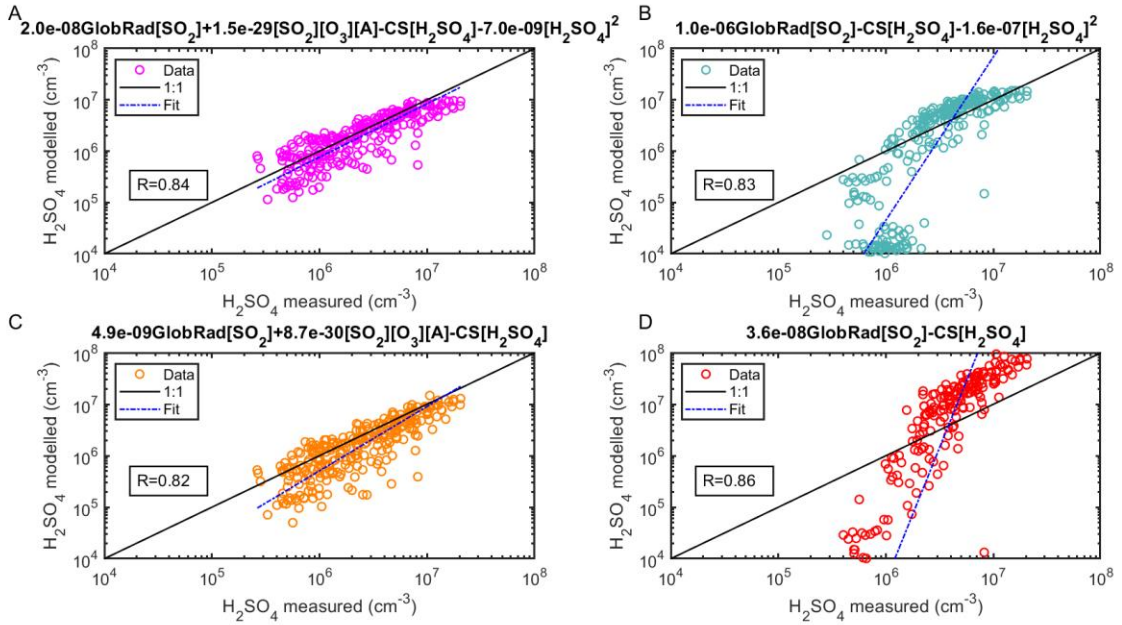


Figure S 22 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at Beijing, China for the testing data set using the four different combinations of source and sink terms. The concentrations shown are 1-hour medians resulting in a total of 268 data points in each subplot. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and the cluster sink term

(Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

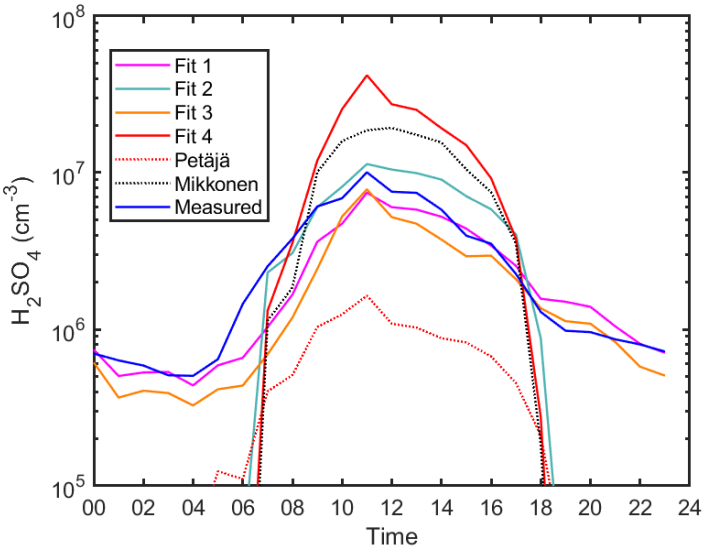


Figure S 23 The diurnal variation of sulphuric acid proxy concentrations using different fits and observed concentrations at in Beijing, China for the testing data set. Median values are shown. Fits 1,2, 3 and 4 corresponds to the Equations 2, 4, 5, and 6, respectively. Petäjä fit shown is applied using the coefficients reported in Petäjä et al. 2009 (Equation 7). Mikkonen fit shown is applied using the coefficients reported in Mikkonen et al. 2011 (Equation 8).

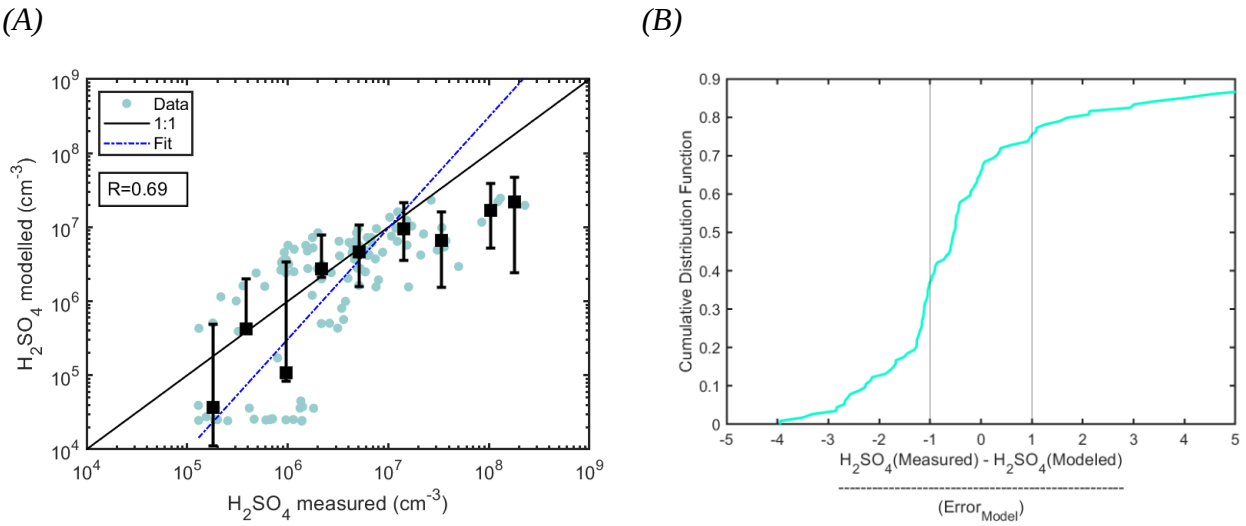


Figure S 24 Sulphuric acid concentrations modelled as a function of measured sulphuric acid at Kilpilahti, Finland. The concentrations shown are 1-hour medians resulting in a total of 114 data points. The modelled concentrations are the median derived using 10,000 k value combinations specific to the the boreal forest location. The colored data points refer to the modelled or predicted concentrations, the dashed blue line refers to the fit (log(y) = a.log(x)+b) of the aforementioned data points. The black squares are the median modelled concentrations in logarithmically spaced measured sulphuric acid bins and their lower and upper whistkers correspond to 25th and 75th percentiles of the predicted concentrations. (B) Cumulative distribution function of the model error weighted difference between measured and modeled H_2SO_4 concentration (using 114 data points).

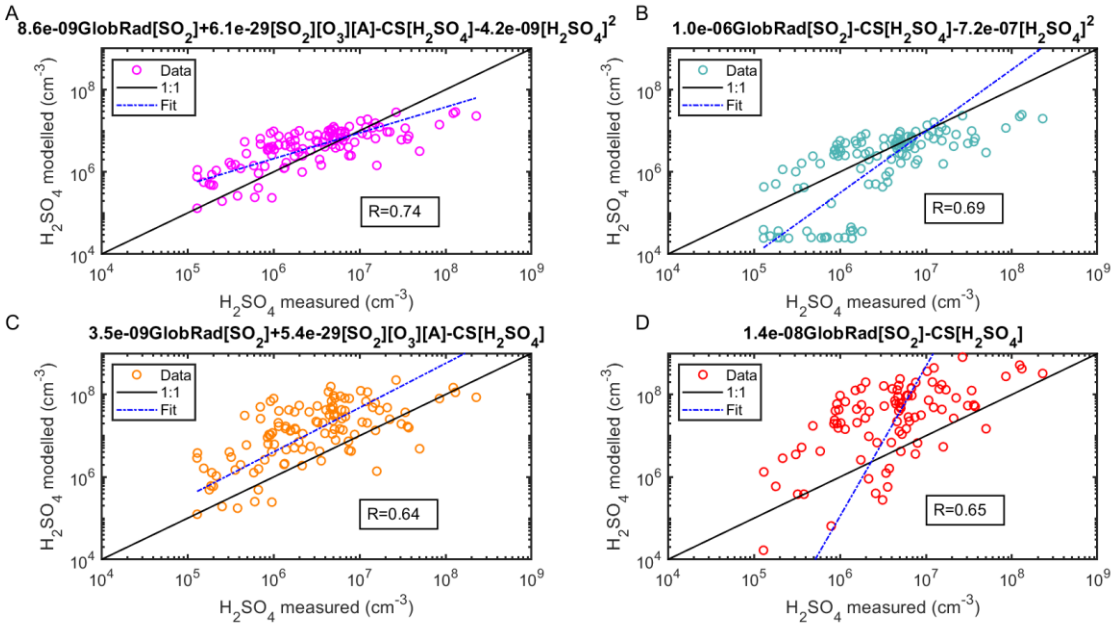


Figure S 25 Sulphuric acid proxy concentration as a function of measured sulphuric acid observed at Kilpilahti, oil refinary Finland using the four different combinations of source and sink terms derived from Hyytiälä. The concentrations shown are 1-hour medians resulting in a total of 114 data points in each subplot. In (A), the full Equation 2 is used, in (B) the equation without the Stabilized Criegee Intermediates source (Equation 4), in (C) the equation without the cluster sink term (Equation 5) and in (D) the equation without both the Stabilized Criegee Intermediates source and

the cluster sink term (Equation 6). The 'Fit' refers to the fitting between the measured and the proxy calculated sulphuric acid concentration (log(y) = a.log(x)+b).

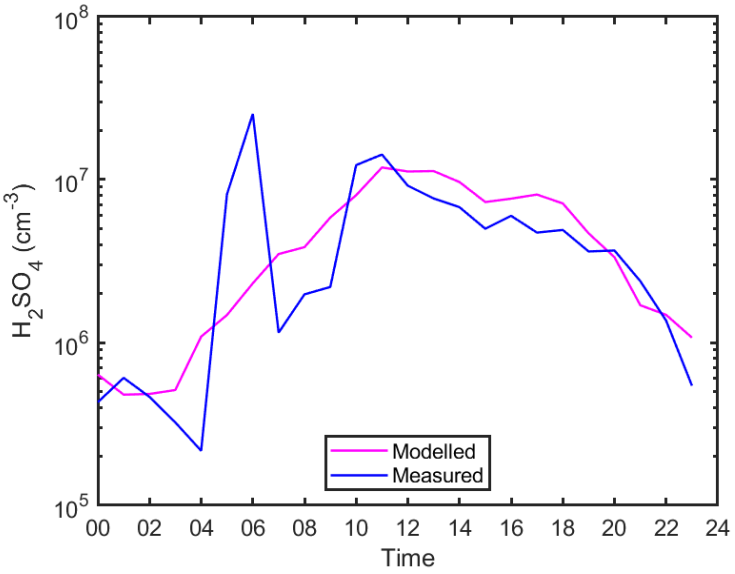


Figure S 26 The diurnal variation of sulphuric acid proxy concentrations observed concentrations at Kilpilahti, industrial area, Finland. Median values are shown. The modelled concentration is predicted using Equation 9 using the k values derived from Hyytiälä SMEAR II station.

311
312
313

References

Laakso, L., Petaja, T., Lehtinen, K. E. J., Kulmala, M., Paatero, J., Horrak, U., Tammet, H., and Joutsensaari, J.: Ion production rate in a boreal forest based on ion, particle and radiation measurements, Atmos Chem Phys, 4, 1933-1943, DOI 10.5194/acp-4-1933-2004, 2004.

McElreath, R.: Statistical rethinking: A Bayesian course with examples in R and Stan, Chapman and Hall/CRC, 2018.

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petaja, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M., and Laaksonen, A.: A statistical proxy for sulphuric acid concentration, Atmos Chem Phys, 11, 11319-11334, 10.5194/acp-11-11319-2011, 2011.

Petäjä, T., Mauldin Iii, R., 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, 10.5194/acp-9-7435-2009, 2009.